Open-Attribute Recognition for Person Retrieval: Finding People Through Distinctive and Novel Attributes

Minjeong Park[1] Hongbeen Park[1] Sangwon Lee[2] Jinkyu Kim[1]
[1]Department of Computer Science and Engineering, Korea University [2]Korea Telecom Research
{minjeongpark, qkrghdqls1, jinkyukim}@korea.ac.kr
lee.sangwon@kt.com

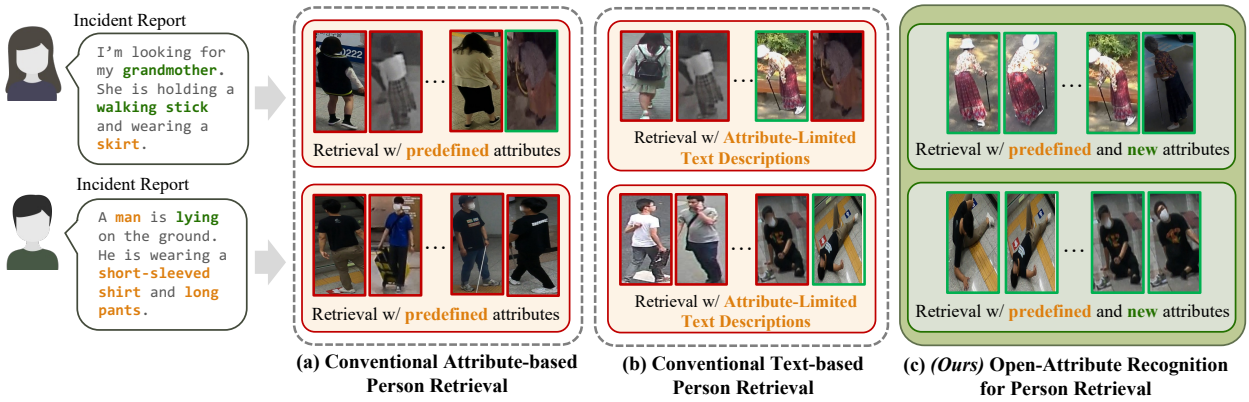


Figure 1. Examples of Open-Attribute Recognition for Person Retrieval in real-world scenarios. Ranking lists are arranged from high to low similarity, with the most similar results appearing on the left. Orange-colored attributes are present in benchmark datasets but typically represent common attributes across individuals, making them less distinctive for person differentiation. In contrast, Green-colored attributes represent distinctive features that facilitate person differentiation but are not present in the benchmarks.

Abstract

*An efficient search for a specific person in the wild requires identifying attributes that distinguish them from others. However, existing person retrieval tasks using attributes either define attribute spaces that are limited mainly to clothing and accessories, making it difficult to distinguish individuals in a truly discriminative way, or adopt a closed-set assumption in which all attributes that appear at test time are also present during training, limiting their applicability in real-world scenarios. To address these challenges, we present a new person retrieval task, **O**pen-**A**ttribute Recognition for **P**erson **R**etrieval (**OAPR**), which aims to retrieve individuals based on attribute cues, regardless of whether those attributes were seen during training. To support this task, we construct an open-attribute dataset by reorganizing four widely used pedestrian attribute datasets (PA-100K, PETA, RAPv1, RAPv2) and evaluation protocols tailored to OAPR. Furthermore, we propose an OAPR framework that learns generalizable body-part representations capable of handling both base and novel attributes. Comprehensive experiments demonstrate the necessity of the OAPR task and the effectiveness of our framework. The source code and pre-trained models will be publicly available upon publication.*

1. Introduction

Person attributes, such as age, gender, hair type, and clothing, play a crucial role in identifying individuals across various computer vision tasks. In particular, attribute-based [10, 12, 32] and text-based person retrieval approaches [8, 33, 40] have leveraged such attributes to retrieve individuals and have achieved remarkable progress. However, such tasks still face significant limitations when deployed in real-world scenarios.

Consider the practical scenario of searching for a missing person or an individual in distress, as illustrated in Fig. 1. In such cases, conventional attribute-based person retrieval (ABPR) approaches assume that all attribute

classes are consistently available during both training and inference. However, this assumption limits their applicability to real-world scenarios where the attributes are absent in the training data. Furthermore, most of the attributes in benchmark datasets are often shared across individuals, making them less distinctive for person differentiation. Specifically, as illustrated on Fig. 2, in RAPv2 [18], the largest dataset in terms of attribute classes among existing pedestrian attribute recognition (PAR) benchmark datasets, the attribute categories are limited to hair type, clothing, attachment, age, gender, body shape, role, and action, and most attributes within these categories represent common appearances among individuals (e.g., long hair, glasses), which limits their utility in identifying individuals with distinctive traits. For example, as shown in Fig. 1(a), when searching for an "elderly woman" wearing a "skirt" and holding a "walking stick", conventional ABPR may only recognize "skirt", however, these are overly common, making it insufficient for distinguishing a target individual (see the first example in Fig. 1(a).). Moreover, even when three predefined attributes are combined as retrieval queries (e.g. combination of male, short-sleeved shirt and long pants), they often describe a large number of individuals, resulting in suboptimal results (see the second example in Fig. 1(a)).

Beyond ABPR, another line of work focuses on text-based person retrieval (TBPR), where natural language descriptions are used to describe the person. While such descriptions may convey richer semantic information, we observe that the annotated descriptions in existing TBPR benchmarks are still limited to clothing and accessory cues as shown in Fig. 3. Consequently, TBPR inherits a similar limitation to ABPR. Moreover, since current TBPR methods are typically trained in an identity-centric manner, when multiple individuals share similar attributes, this ID-driven objective forces the model to separate them in the embedding space. For example, as shown in the third and fourth samples in Fig. 1(b), even an image that visually matches the given text description is treated as a negative sample if it does not correspond to the annotated identity.

To address these limitations, in this paper, we propose an **O**pen-**A**ttribute Recognition for **P**erson **R**etrieval, dubbed as OAPR, a new person retrieval approach, and formulate the task as text-to-image search. As shown in Fig. 1(c), the task aims to retrieve individuals given the attribute cues, regardless of whether those attributes were trained. The model is trained on a set of base attributes and evaluated on both base (seen) and novel (unseen) attributes in the test set. Under this formulation, if a model trained only on clothing-related attributes is required to recognize attributes across diverse categories, it may struggle to effectively leverage its learned representations, leading to poor generalization performance. To support robust prediction of novel attributes, we define a base attribute set that enables

Predefined Attribute Classes on RAPv2[18]

hs-BaldHead	hs-LongHair	hs-BlackHair
hs-Hat	hs-Glasses	ub-Shirt
ub-Sweater	ub-Vest	ub-TShirt
ub-Cotton	ub-Jacket	ub-SuitUp
ub-Tight	ub-ShortSleeve	ub-Others
lb-LongTrousers	lb-Skirt	lb-ShortSkirt
lb-Dress	lb-Jeans	lb-TightTrousers
shoes-Leather	shoes-Sports	shoes-Boots
shoes-Cloth	shoes-Casual	shoes-Other
attachment-Backpack	attachment-ShoulderBag	attachment-HandBag
attachment-Box	attachment-PlasticBag	attachment-PaperBag
attachment-HandTrunk	attachment-Other	AgeLess16
Age17-30	Age31-45	Age46-60
Female	BodyFat	BodyNormal
BodyThin	Customer	Employee
action-Calling	action-Talking	action-Gathering
action-Holding	action-Pushing	action-Pulling
action-CarryingByArm	action-CarryingByHand	action-Other

Figure 2. Attributes in RAPv2. Each color represents a different category. (e.g. hair type, clothing, attachment, age, gender, body shape, role and action.)

Figure 3. Attribute words in existing text-based person retrieval datasets.

effective knowledge transfer from base attributes to novel ones. To this end, we present an Open-Attribute dataset by rebuilding the widely used PAR benchmark datasets, such as PA-100K [23], PETA [5], RAPv1 [17], and RAPv2 [18], to learn attributes across multiple categories and evaluate the acquired knowledge by inferring novel classes.

To tackle this challenge, we propose a novel framework built upon CLIP [27], leveraging its superior trade-off between performance and efficiency, thereby facilitating practical real-world applications. While CLIP provides a strong foundation for open-attribute recognition, prior work has shown that it struggles to capture fine-grained semantic regions [27], which conflicts with the characteristics of person attributes. To address this limitation, our framework focuses on learning generalizable representations across a wide range of attribute categories. Specifically, we extract pseudo body features and leverage them to optimize the learnable body prompts. Subsequently, body prompt features are selected for each attribute and used to determine its presence or absence.

To sum up, the main contributions of this paper are as the following four aspects:

- We propose a new person retrieval approach, Open-Attribute Recognition for Person Retrieval (**OAPR**), which aims to retrieve individuals based on given attributes, regardless of whether those attributes were seen during training.
- We present Open-Attribute dataset by rebuilding four widely-used datasets for open-attribute recognition. This enables the model to learn across diverse attribute domains and facilitates its evaluation of knowledge transfer.
- We introduce a novel framework specifically designed for OAPR. Our framework aims to learn generalizable information covering a wide range of attribute categories.
- Comprehensive experiments on 4 datasets demonstrate that our framework achieves superior performance of retrieving people based on base and novel attributes. We hope our work inspires further research on OAPR, which is more practical for real-world applications.

2. Related Work

Pedestrian Attribute Recognition (PAR) The pedestrian attribute recognition task aims to identify various detailed attributes of an individual given an image. Most PAR methods follow a closed-world setting, assuming attribute classes remain identical across training and testing phases. VTB [4] formulates PAR as a multimodal multi-label classification problem using BERT [6] and ViT [7]. Building on this, PromptPAR [29] leverages CLIP [27] with region-aware prompt learning, while ViTA-PAR [24] extends this idea with attribute-specific prompts that capture a wider range of attribute semantics. Although effective on seen attributes, their closed-set assumption limits generalization to novel attributes and reduces real-world practicality. POAR [34] formulates pedestrian open-attribute recognition as an image-to-text search task, with a primary focus on addressing domain shift. In contrast, our OAPR task emphasizes the ability to infer novel attributes that have never appeared during training. Moreover, we formulate OAPR as a text-to-image search task.

Person Retrieval with Attributes Attributes have been utilized as auxiliary information for person retrieval. Among related tasks, text-based person retrieval (TBPR) primarily relies on natural language descriptions to identify individuals [2, 8, 9, 15, 20, 30]. While TBPR motivates the use of attributes for retrieval, its primary goal is to retrieve the unique identity associated with the given text description, which differs from the objective of OAPR. Moreover, OAPR directly takes attribute classes as input. Beyond TBPR, attribute-based person retrieval (ABPR) [10, 12, 32] also leverage attributes for retrieval, but they are limited to predefined attribute classes within a dataset. In contrast, OAPR explore retrieving individuals based on novel attributes.

Visual Language Models (VLM) and Prompt Learning (PL) Recently, pre-trained vision-language models (VLMs) [13, 19, 27] show significant performance gain on various downstream tasks [21, 38, 39]. CLIP [27] learns robust image–text representations from large-scale natural language supervision, enabling strong zero-shot transfer on diverse vision benchmarks. BLIP [19] introduces a unified vision–language pre-training framework that leverages both web and curated captions to improve image–text understanding and generation. ALIGN [13] scales contrastive image–text pre-training to billions of noisy image–text pairs, achieving competitive zero-shot performance without task-specific fine-tuning. We adopt CLIP as our baseline to leverage its strong performance and computational efficiency, enabling practical deployment in real-world applications.

Prompt learning (PL) has emerged as a parameter-efficient strategy that achieves competitive performance without the need for full network fine-tuning [3, 11, 22, 31, 34]. CoOp [36] introduces learnable text prompts that help adapt VLM to downstream tasks, and shows promising results on open-vocabulary learning tasks [31]. CoCoOp[35] is the extended version of CoOp that learns a light-weight visual network to provide meta tokens for each image. MaPLe [16] introduces multi-modal prompt learning by jointly optimizing prompts in both the vision and language branches with a cross-modal coupling function. Building on these tasks, we introduce learnable body and text prompts to enhance visual and attribute representations.

3. Open-Attribute Recognition for Person Retrieval

3.1. Task Definition

We treat the open-attribute recognition for person retrieval (OAPR) as text-to-image search. Regardless of whether an attribute was trained or not, the goal of this task is to retrieve individuals based on a given attribute. Specifically, given the attributes of target person $\pi = \{a_1, a_2, \cdots, a_r\}, a_i \in \mathcal{A}$, the model need to find the K person images $\mathcal{X} = \{X_1, , X_2, \cdots, X_K\}$. Here, $\mathcal{A}$ defines the attribute set and $\mathcal{A} = \{\mathcal{A}_{base}, \mathcal{A}_{novel}\}, \quad \mathcal{A}_{base} \cap \mathcal{A}_{novel} = \varnothing$. The model is trained on the base attribute space ($\mathcal{A}_{base}$) and evaluated on both the base and novel attribute spaces ($\mathcal{A}_{base} \cup \mathcal{A}_{novel}$).

3.2. Open-Attribute Dataset

In this section, we provide a detailed explanation of the dataset reconstruction process for the OAPR task. We split a base (for training) and novel (for test) attributes on the widely used attribute recognition datasets, such as PA-100K [23], PETA [5], RAPv1 [17] and RAPv2 [18]. We leverage attribute text similarity and hierarchical relationships to ensure effective knowledge transfer from base attributes to novel attributes. The overall process is shown in Fig. 4.

Step 1. Attribute Filtering and Verbalization Since some existing attributes are too ambiguous for effective person search, we filter out those that may not provide meaningful discriminative cues for person search. For instance, *"attachment-Other"* in RAPv2 encompasses a wide range of possibilities, making it difficult to associate with a specific target in the retrieval process. In addition, predefined attribute classes are typically represented in a structured, compact format, such as *upperBodyShortSleeve* and *attachHandBag*, which is not directly suitable for text encoder inputs. To address this, we reformulate each attribute into a natural language phrase. Next, we convert each attribute into a natural language description. For example, the attribute classes of *"upperBodyShortSleeve"* is converted to *"Wearing short-sleeve upper body clothing"* and *"attachHandBag"* is changed into *"Carrying a handbag"*. Finally, we can get 26 attributes in PA-100K, 33 attributes in PETA, 50 attributes in RAPv1 and RAPv2.

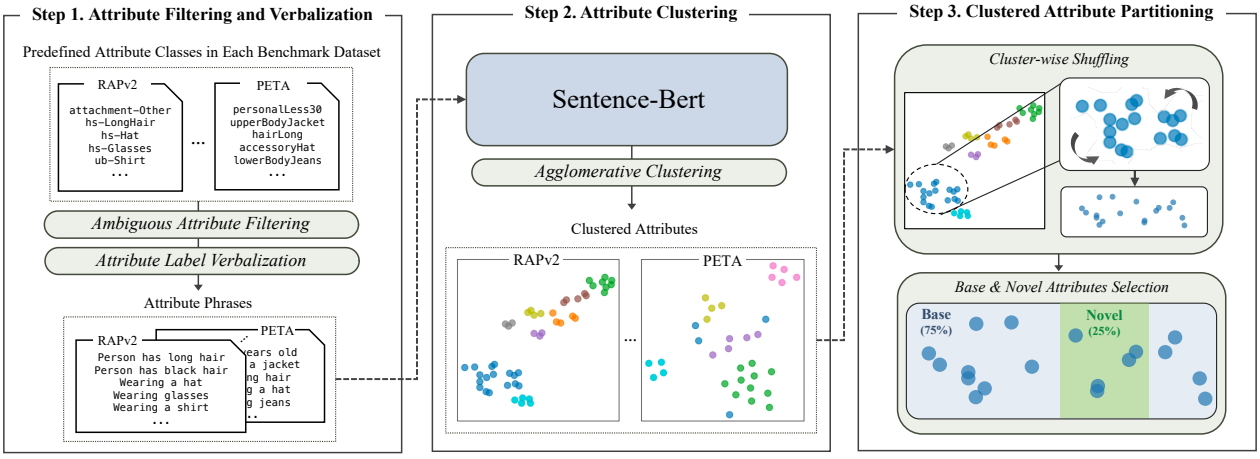


Figure 4. Illustration of the preprocessing pipeline for constructing the Open-Attribute Dataset, comprising three stages: (1) attribute filtering and verbalization, (2) attribute clustering, and (3) clustered attribute partitioning.

Step 2. Attribute Clustering Manually splitting base and novel attributes can be time-consuming and subjective; therefore, we propose attribute clustering based on text features. We input the phrase into Sentence-Bert [28] to encode each attribute phrase into a vector, and employ Agglomerative Clustering [26] to cluster the vectors, leveraging the hierarchical structure of attributes (e.g. grouping clothing-related features into subcategories). Note that we use Sentence-BERT instead of CLIP text encoder to avoid information leakage, as CLIP is already utilized in our model. We determine the number of clusters ($|\mathcal{C}|$) based on the characteristics of each dataset. For instance, the attributes in the PETA [5] primarily pertain to appearance-related factors, such as clothing, accessories, age, and bags. In contrast, the RAPv2 [18] dataset extends beyond these attributes to include action-related attributes (e.g., making a phone call) and body shape information (e.g., slimness). Based on this observation, we assign 7 clusters to PA-100K, 6 clusters to PETA, 8 clusters to RAPv1 and RAPv2.

Table 1. Statistics of Open-Attribute Dataset.

Dataset	#. of Clusters ($\|\mathcal{C}\|$)	Attributes		Train	Test
		Base	Novel		
PA-100K [23]	7	18	8	90,000	10,000
PETA [5]	6	23	10	11,400	7,600
RAPv1 [5]	8	35	16	33,268	8,317
RAPv2 [5]	8	38	16	67,943	16,985

Step 3. Clustered Attribute Partitioning To ensure that our model can effectively transfer from base attributes to novel attributes, we partition each cluster into base and novel classes. Specifically, we randomly shuffle the attributes within each cluster and designate 25% of attributes from each cluster as novel attributes, while the remaining attributes serve as base attributes. Finally, we can get 18 base and 8 novel classes on PA-100K, 23 base and 10 novel classes on PETA, 35 base and 16 novel classes on RAPv1, and 38 base and 16 novel classes on RAPv2. For dataset partitioning, we follow the standard protocol [14]. The statistics of the open-attribute dataset are presented in Tab. 1, with more detailed information available in the supplementary materials.

3.3. Evaluation Metrics

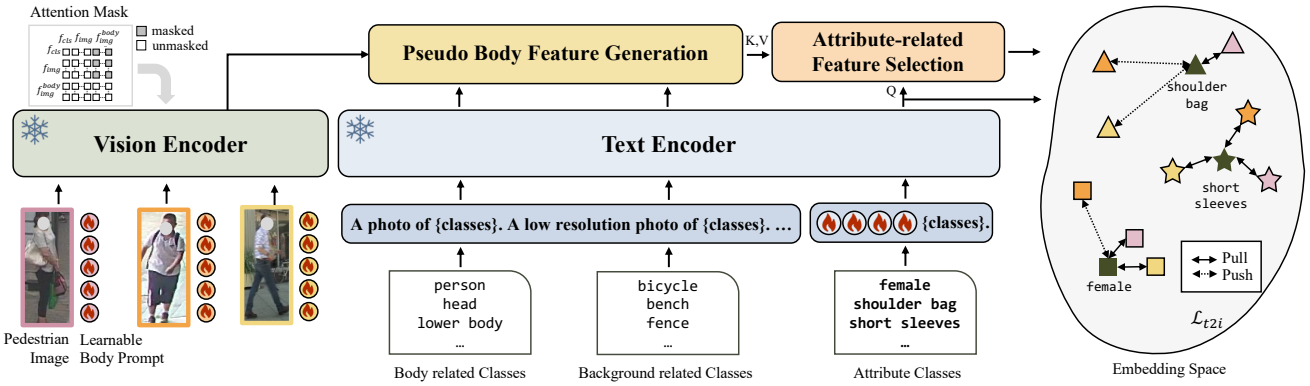
To evaluate the person retrieval performance, we introduce two metrics. (i) Precision@K for label (P@K-lbl) measures the average correctness of individual attribute conditions within the top-K retrieved images. Specifically, we compute the proportion of correctly matched attribute labels per attribute and average this score across all selected attributes and all queries. It indicates how accurately the retrieved individuals match the given attributes. (ii) Precision@K for instance (P@K-ins) considers a retrieval successful only when all attribute conditions in the query are simultaneously satisfied by K retrieved images. For each query, we check whether at least one of the top-K images perfectly matches the entire attribute set. The final metric is computed as the average success rate across all queries. More detailed information is available in the supplementary materials.

4. Proposed Approach

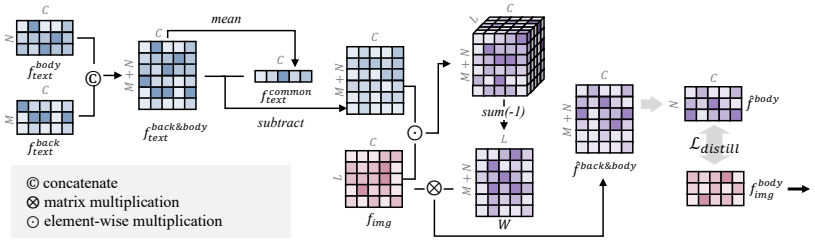
4.1. Approach Overview

As shown in Fig. 5, our framework consists of a CLIP-based vision encoder, a text encoder, a pseudo-body feature generation module, and an attribute-related feature selection module. Note that since real-time performance is crucial for this task, we adopt CLIP instead of computationally heavy multimodal large language models.

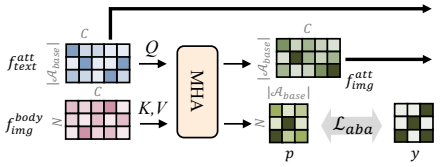
For the vision encoder, inspired by prior studies showing that diagonally prominent attention maps enhance local visual semantics by reducing patch-level noise [22, 37],



(a) Overview of the training process of our proposed framework.



(b) The process of the proposed pseudo body feature generation.



(c) The process of the proposed attribute-related feature selection.

Figure 5. Our framework consists of the CLIP vision and text encoders, a pseudo body feature generation module, and an attribute-related feature selection module. In the embedding space, the dark green triangle denotes the attribute text features f_{text}^{att}. Identical colors represent the same individual, and identical shapes denote the same attribute category.

we replace the V Q-K self-attention blocks with V-V self-attention in the vision encoder. Additionally, we apply a masking strategy to prevent the learnable prompts from interfering with the pre-trained representations. The learnable body prompt $Z = [z_1, z_2, ..., z_N]$ and an input image X are processed through the vision encoder to obtain the $f^{body} \in \mathbb{R}^{N \times C}$, and $f_{cls} \in \mathbb{R}^{C}$ and $f_{img} \in \mathbb{R}^{L \times C}$ derived from Z and X, respectively.

$$[f_{cls}, f_{img}, f_{img}^{body}] = \text{VisEnc}([X, Z]), \tag{1}$$

For textual attribute embeddings, we define N body-related classes (e.g., lower body, head) and M background-related classes (e.g., bicycle, bench). We employ the prompt ensemble strategy [27], which combines various context templates (e.g., A photo of class, A low-resolution photo of class) to construct body and background prompts. By passing these prompts through the text encoder, we obtain the body and background text features $f_{text}^{body} \in \mathbb{R}^{N \times C}$ and $f_{text}^{back} \in \mathbb{R}^{M \times C}$. We adopt prompt learning inspired by CoOp [36] for $|\mathcal{A}_{base}|$ attribute classes, appending a learnable prompt to the attribute embeddings, and obtain the attribute text feature $f_{text}^{att} = [f_{text}^{att(1)}, ..., f_{text}^{att(|\mathcal{A}_{base}|)}] \in \mathbb{R}^{|\mathcal{A}_{base}| \times C}$. Then f_{img}, f_{text}^{body} and f_{text}^{back} pass through the pseudo body feature generation module to produce the pseudo body feature, which is distilled into f_{img}^{body}.

Subsequently, the proposed attribute-related feature selection module takes f_{text}^{att} as the query and f_{img}^{body} as the key and value, deriving $f^{att} \in \mathbb{R}^{|\mathcal{A}_{base}| \times C}$ with enhanced attribute-specific visual representations. For effective training, we introduce an attribute part association loss to enhance the model's ability to learn the body parts associated with each attribute. Finally, as shown in Fig. 5a, we compute the similarity between the batch of attribute-conditioned visual features and the text features of each attribute, pulling positive pairs closer and pushing negative pairs apart via a text-to-image contrastive loss.

Instead of memorizing individual attribute labels, our model learns body-part–conditioned representations that capture generic visual patterns shared across various attributes. By decoupling attribute semantics from specific categories and grounding them in a set of reusable body-part representations, the model can transfer these representations to novel attributes that were never seen during training. We provide a more detailed explanation in the following sections.

4.2. Pseudo Body Feature Generation

Since regional body information is crucial for attribute recognition, many researchers suggest various approaches to obtain body part features [1, 24, 29, 34]. Different from previous studies, the proposed pseudo body feature generation leverages the text information to obtain a robust body

part feature, as shown in Fig. 5b. To eliminate the background disturbance and obtain the body part features, we compute the common feature across the classes and remove it. Specifically, we concatenate f_{text}^{back} and f_{text}^{body} to obtain $f_{text}^{back\&body}$, and calculate the mean feature vector f_{text}^{common}, which represents the class-irrelevant components common to both body and background. By subtracting this common representation, the remaining features emphasize class-specific directions, thereby enhancing the separability between body and background features. We perform an element-wise multiplication with f_{img} and sum the features along the channel dimension to obtain W, which highlights body and background regions with improved discrimination. Finally, we can obtain the $\hat{f}^{back\&body}$ through a matrix multiplication between f_{img} and W as follows:

$$W = {f_{img}}^{T}(f_{text}^{back\&body} - f_{text}^{common}) \tag{2}$$

$$\hat{f}^{back\&body} = W f_{img} \tag{3}$$

We select only body part proportion $\hat{f}^{body}$ from $\hat{f}^{back\&body}$, which serves as the pseudo body feature. The pseudo feature supervises f_{img}^{body} via an L2 loss to distill its information:

$$\mathcal{L}_{distill} = ||f_{img}^{body} - \hat{f}^{body}||_2^2 \tag{4}$$

4.3. Attribute-Related Feature Selection

Given that certain attributes exhibit spatial dependencies across multiple body parts (e.g., long hair extends across the head and upper body, and a dress spans both upper and lower body), we introduce an attribute-related feature selection module to adaptively focus on the most relevant regions for each attribute through an attention mechanism, as shown in Fig. 5c. Specifically, given that f_{text}^{att} and f_{img}^{body}, attention weight $p \in \mathbb{R}^{N \times |\mathcal{A}_{base}|}$ and the attribute-conditioned visual features $f_{img}^{att} = [f_{img}^{att(1)}, ..., f_{img}^{att(|\mathcal{A}_{base}|)}] \in \mathbb{R}^{|\mathcal{A}_{base}| \times C}$ computed as:

$$p = \text{Softmax}(\frac{f_{text}^{att} W^Q (f_{img}^{body} W^K)^\top}{\sqrt{d_k}}) \tag{5}$$

$$f_{img}^{att} = p(f_{img}^{att} W^V) \tag{6}$$

where W^Q, W^K, W^V represent the query, key, and value projection matrices, respectively, incorporating the multi-head mechanism.

At the early training stage, we observed that f_{img}^{body} exhibits highly similar features across different images and attributes, which hinders effective learning and convergence of the model. To address this issue, we propose an attribute-body association (ABA) loss that provides supervision on the body part relevant to each attribute. Specifically, it encourages each attribute to align with its corresponding body part; for example, the attribute "short-sleeved t-shirt" is associated with the upper-body region. The ABA loss is formulated as:

$$\mathcal{L}_{aba} = \frac{1}{|\mathcal{A}_{base}|} \sum_{i=1}^{|\mathcal{A}_{base}|} \sum_{j=1}^{N} -y_j^{(i)} \log p_j^{(i)} \tag{7}$$

where $y_j^{(i)}$ indicates whether the i-th attribute belongs to the j-th body part.

Finally, we calculate the similarity between the batch of the attribute-conditioned visual features and the text attribute feature. To achieve this, we propose a text-to-image contrastive loss to effectively align image features with text features, ensuring a more precise and semantically meaningful feature correspondence.

$$\mathcal{L}_{t2i} = \sum_{i=1}^{|\mathcal{A}_{base}|} -\log\left(\frac{S_B^{i+}}{S_B^{i+} + w_{neg} S_B^{i-}}\right) \tag{8}$$

$$S_B^{i+} = \exp(I_B^{i+} f_{text}^{att(i)}/\tau),\ S_B^{i-} = \exp(I_B^{i-} f_{text}^{att(i)}/\tau) \tag{9}$$

where $f_{text}^{att(i)}$ refers to the i-th text attribute feature. I_B^{i+} denotes the set of positive pairs between the batch of i-th attribute-conditioned visual features $f_{img}^{att(i)}$ and the corresponding $f_{text}^{att(i)}$, while I_B^{i-} denotes the set of negative pairs.

Overall, during the training phase, we employ three loss functions for the total loss function given:

$$\mathcal{L}_{train} = \mathcal{L}_{t2i} + \lambda_{distill}\mathcal{L}_{distill} + \lambda_{aba}\mathcal{L}_{aba} \tag{10}$$

where $\lambda_{distill}, \lambda_{aba}$ are the hyperparameters.

5. Experiments

Implementation Details We adopt the image and text encoders from pre-trained CLIP (ViT-B/16) [27] and freeze both image and text encoders. The proposed model is implemented with PyTorch [25] and trained on an A6000 GPU. The number of distinct body features (N) is set to 5, and we use the following body part classes: *"person"*, *"head"*, *"upper body"*, *"lower body"*, *"holding something"*. The number of the learnable prompt is set to 66 tokens, and the prompt is shared across all attributes. V is set to 6. The learning rate for the learnable text prompt is set to 0.005, while the learning rate for the cross-attention module is 0.001. The model is trained for 100 epochs using a cosine scheduler. The temperature parameter τ is set to 0.07, and the weighting factor for negative samples, w_{neg}, is set to 50. The hyperparameters $\lambda_{distill}$ and λ_{aba} are set to 0.4 and

Table 2. Results of Attribute-to-Person Retrieval in P@K-lbl and P@K-ins. The model marked with * is the reimplemented model, and the **best** and second-best results are shown in **bold** and underlined, respectively. Improvements over the second-best are highlighted in red. VLM, PL, and PAR denote visual–language models, prompt-learning methods, and pedestrian attribute recognition methods, respectively.

	Datasets / Models	PA-100K [23]				PETA [5]				RAPv1 [17]				RAPv2 [18]			
		Base		Novel		Base		Novel		Base		Novel		Base		Novel	
		P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl
VLM	CLIP[27]	50.41	50.10	50.89	49.73	50.21	50.05	50.00	49.83	50.08	50.01	50.00	50.02	50.07	50.04	50.21	50.17
VLM	BLIP[19]	49.92	50.59	50.00	50.80	50.05	50.00	50.00	50.61	50.04	50.01	49.90	50.02	49.96	50.01	49.90	50.42
VLM	ALIGN[13]	47.71	50.13	49.11	46.96	50.20	50.15	44.44	48.72	50.63	49.87	48.12	49.46	49.84	49.90	49.69	50.96
PL	CoOp[36]	51.51	50.02	51.34	50.27	50.20	50.10	54.17	51.22	49.98	49.99	50.10	49.98	50.02	50.01	49.79	50.02
PL	CoCoOp[35]	50.57	50.72	50.89	50.54	50.10	50.02	50.00	50.17	50.55	50.05	47.60	49.58	50.11	50.05	50.00	50.12
PL	MaPLe[16]	51.22	50.13	51.39	50.98	50.19	50.45	54.72	51.44	50.02	50.01	50.41	50.14	49.98	50.02	50.21	50.02
PAR	VTB*[4]	50.00	49.92	53.57	50.45	50.39	50.50	43.33	49.50	49.87	49.61	50.94	50.10	50.02	50.00	50.00	49.27
PAR	PromptPAR*[29]	50.57	50.55	53.12	50.98	50.29	50.12	56.11	51.50	50.18	50.10	51.66	50.25	50.03	50.06	51.45	50.29
PAR	ViTA-PAR*[24]	51.22	51.65	51.78	49.73	51.28	51.24	52.77	51.94	50.23	50.17	50.31	50.29	50.39	50.24	50.45	50.50
PAR	POAR*[34]	51.51	50.72	53.21	50.45	50.44	50.22	53.61	52.83	49.72	49.50	50.08	50.05	50.04	50.05	50.31	50.12
	Ours	**54.00** (+2.49)	**52.30** (+0.65)	**55.35** (+1.78)	**52.67** (+1.69)	**52.66** (+1.38)	**51.39** (+0.15)	**59.16** (+3.05)	**55.27** (+2.44)	**50.65** (+0.10)	**50.46** (+0.29)	**53.85** (+2.19)	**51.37** (+1.08)	**50.51** (+0.12)	**50.60** (+0.36)	**52.18** (+0.73)	**51.29** (+0.69)
		P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl	P@1-lbl	P@5-lbl
VLM	CLIP [27]	24.84	25.10	25.89	25.00	25.43	25.06	25.00	24.67	25.04	24.99	25.00	24.88	25.07	25.00	25.21	25.13
VLM	BLIP[19]	25.33	25.46	25.00	26.61	25.10	25.06	25.00	25.33	25.00	25.03	25.00	24.88	24.89	25.00	24.79	25.13
VLM	ALIGN[13]	23.20	25.20	22.32	23.93	24.41	25.10	21.11	23.44	25.13	24.97	23.33	24.54	24.82	24.94	26.04	25.25
PL	CoOp [36]	25.82	25.07	25.89	25.00	25.30	25.10	30.00	26.00	25.00	24.99	25.00	25.08	25.00	25.01	24.79	25.00
PL	CoCoOp [35]	25.16	25.56	25.00	25.00	25.10	25.02	25.00	25.33	25.21	25.03	22.92	24.63	25.07	25.07	25.21	24.88
PL	MaPLe[16]	25.16	25.03	26.78	25.35	25.49	25.15	28.88	27.44	25.40	25.01	24.95	25.25	25.03	25.01	25.41	25.16
PAR	VTB* [4]	25.49	24.15	27.67	26.43	25.36	25.60	18.33	23.67	24.83	24.64	26.87	24.67	25.04	25.00	24.79	24.71
PAR	PromptPAR*[29]	25.98	25.55	27.67	26.07	25.19	25.07	31.11	27.44	25.16	25.07	25.83	25.33	25.03	25.06	26.25	25.58
PAR	ViTA-PAR*[24]	25.98	26.34	27.67	25.53	25.77	26.08	27.11	27.77	25.25	25.16	25.41	25.37	25.28	25.27	26.45	25.25
PAR	POAR* [34]	25.82	25.07	25.89	25.00	25.49	25.19	28.88	28.11	25.04	25.03	25.00	25.08	25.07	25.28	26.45	25.29
	Ours	**28.43** (+2.45)	**26.96** (+0.62)	**29.46** (+1.79)	**28.03** (+1.60)	**27.37** (+1.88)	**26.30** (+0.70)	**33.88** (+2.77)	**30.77** (+2.66)	**25.58** (+0.33)	**25.42** (+0.26)	**28.33** (+1.46)	**26.37** (+1.00)	**25.35** (+0.07)	**25.33** (+0.05)	**27.91** (+1.45)	**27.20** (+0.62)

Table 3. Ablation studies.

(a) Different Loss Combinations

Model	$\mathcal{L}_{t2i}$	$\mathcal{L}_{distill}$	$\mathcal{L}_{aba}$	Base (P@1-lbl)	Base (P@5-lbl)	Novel (P@1-lbl)	Novel (P@5-lbl)
A	✓	✗	✗	50.84	50.88	52.77	50.83
B	✓	✓	✗	50.54	50.65	53.33	51.61
C	✓	✗	✓	50.29	51.07	54.72	53.05
Ours	✓	✓	✓	**52.66**	**51.39**	**59.16**	**55.27**
Model	$\mathcal{L}_{t2i}$	$\mathcal{L}_{distill}$	$\mathcal{L}_{aba}$	Base (P@1-ins)	Base (P@5-ins)	Novel (P@1-ins)	Novel (P@5-ins)
A	✓	✗	✗	25.79	25.75	26.66	25.88
B	✓	✓	✗	25.49	25.83	28.33	26.44
C	✓	✗	✓	25.29	25.87	30.55	28.55
Ours	✓	✓	✓	**27.37**	**26.30**	**33.88**	**30.77**

(b) Different Body Part Classes Combinations

Model	Body Part Types	Base (P@1-lbl)	Base (P@5-lbl)	Novel (P@1-lbl)	Novel (P@5-lbl)
a	person, upper body, lower body	52.02	50.97	54.16	51.18
b	person, face, hair, upper body, lower body, holding something	52.52	51.12	55.83	52.16
Ours	person, upper body, lower body, holding something	**52.66** (+0.14)	**51.39** (+0.27)	**59.16** (+3.33)	**55.27** (+3.11)
Model	Body Part Types	Base (P@1-ins)	Base (P@5-ins)	Novel (P@1-ins)	Novel (P@5-ins)
a	person, upper body, lower body	26.58	25.75	29.44	25.81
b	person, face, hair, upper body, lower body, holding something	26.68	25.59	31.67	26.88
Ours	person, upper body, lower body, holding something	**27.37** (+0.69)	**26.30** (+0.55)	**33.88** (+2.21)	**30.77** (+3.89)

0.1, respectively. Further configuration details are provided in the supplementary material.

Quantitative Results We report the comparison with vision-language models (CLIP [27], BLIP [19], ALIGN [13]), prompt learning methods (CoOp [36], CoCoOp [35], MaPLe [16]), and pedestrian attribute recognition models, including closed-set (VTB [4], PromptPAR [29], ViTA-PAR [24]) and open-set (POAR [34]) approaches. All methods are trained with base attributes from the train set, and their performance is reported on the test set. Performance is reported on all two-attribute combinations ($|\pi| = 2$). For a fair comparison, all models, except VLMs, are evaluated with CLIP ViT-B/16, and the best-performing model on novel attributes during validation is selected. As shown in Tab. 2, our framework consistently achieves superior attribute-to-person retrieval performance across all datasets, demonstrating strong performance on both base and novel attributes. Notably, the performance gains become substantially larger for novel attributes. This confirms that our learned body-part–conditioned representations are more effective for transferring knowledge from base to unseen attribute concepts.

Qualitative Results We visualize person retrieval results for both base and novel attributes as shown in Fig. 6. We select novel attributes that are entirely new and not included in any benchmark datasets to evaluate the model's generalization to unseen attribute queries. In addition, to assess cross-domain capability, we use the model trained on RAPv2 [18] to infer samples from the PETA [5], making the evaluation more aligned with real-world scenarios. As shown, the model successfully retrieves individuals conditioned on various untrained attribute categories, such as age (elderly), object (bicycle, stroller), and action (sitting down). These results demonstrate the robustness and adaptability of our approach, confirming its effectiveness in tackling the challenges of the OAPR task.

Effect of Pseudo Body Feature Generation Method As shown in Fig. 7, we visualize the activation map for the "person", "head", "upper body", "lower body" and "holding something" of W in the pseudo body feature genera-

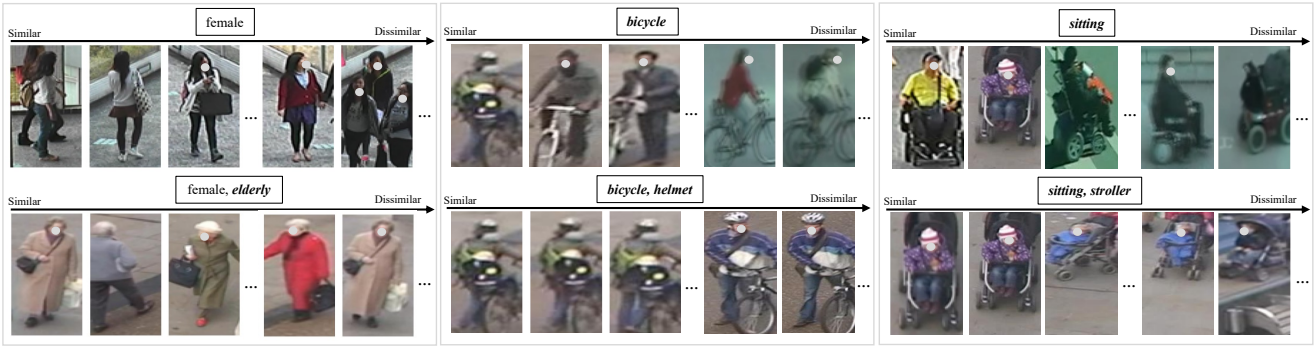


Figure 6. Person retrieval results according to the given attributes. ***Attribute names in bold italic*** indicate ***novel attributes*** that were not seen during training. Our model demonstrates strong retrieval performance across both base and novel attributes.

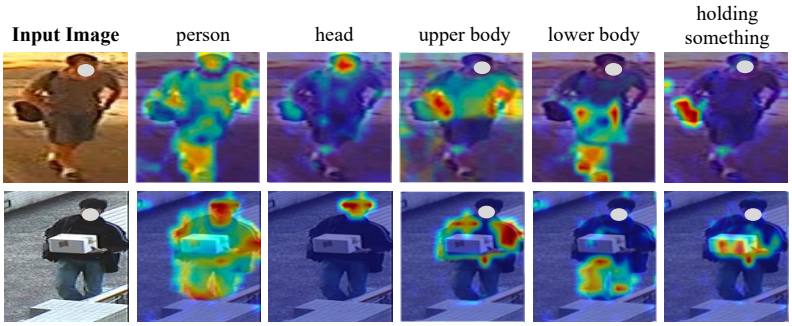


Figure 7. Visualization demonstrating the effectiveness of the pseudo-body generation module.

Table 4. Computational cost comparison.

Models	# of Params.	Inference Time
CLIP [27]	86.19M	35.01ms
BLIP [19]	223.45M	117.23ms
ALIGN [13]	172.12M	93.23ms
VTB [4]	236.56M	39.44ms
PromptPAR [29]	159.09M	32.86ms
ViTA-PAR [24]	134.67M	29.41ms
POAR [34]	107.10M	42.12ms
Ours	125.41M	35.29ms

tion module to demonstrate its effectiveness. The maps are color-coded, where red indicates a high response. As illustrated, our proposed module effectively highlights the regions of interest for each body part without training cost, even under challenging conditions.

Effect of the Proposed Loss We validate the effectiveness of each loss term with an ablation study on PETA [5], as shown in Tab. 3(a). Model A serves as the baseline with only $\mathcal{L}_{t2i}$ for basic text-to-image retrieval. Adding $\mathcal{L}_{distill}$ in Model B distills pseudo-body features into the learnable parameters, while Model C leverages $\mathcal{L}_{aba}$ to supervise the spatial association between each attribute and its corresponding body region. When all three losses are combined, our model achieves the best performance, demonstrating that each component contributes meaningfully to the overall framework.

Analysis of different body part combinations We present combinations of body-part classes to investigate the effect of part granularity on recognition performance on PETA [5], as shown in Tab. 3(b). Across different choices of the number and type of body parts, we observe that using a set of five body parts (person, upper body, lower body, holding something) provides a better trade-off between granularity and robustness. In contrast, adding highly localized parts such as face and hair introduces noisy and overlapping supervision under low-resolution surveillance settings, ultimately leading to worse generalization on novel attributes.

Analysis on Computational Cost Since computational efficiency is crucial for deploying the OAPR task in real-world scenarios, we compare our proposed framework with other PAR models. We calculate the model's parameters, inference time, and inference GFLOPs. We experimented on 64 samples using the same single RTX A6000 GPU. As shown in Tab. 4, our model demonstrates lightweight design and efficient inference, enabling practical deployment in real-world scenarios compared to existing approaches.

6. Conclusion

In this work, we introduced Open-Attribute Recognition for Person Retrieval (OAPR), a new task that aims to retrieve individuals based on attribute cues without assuming that all attributes are known during training. To support this setting, we reconstructed four widely used pedestrian attribute datasets into an open-attribute benchmark. We further proposed an OAPR framework that learns generalizable, body-part–conditioned representations, driven by pseudo-body feature generation and attribute-related feature selection. Comprehensive experiments across four datasets demonstrate that our approach consistently outperforms vision–language, prompt-learning, and PAR baselines on both base and novel attributes. We hope this work inspires further exploration into open-attribute representations and practical retrieval systems capable of handling the vast diversity of real-world attribute descriptions.

References

[1] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15050–15061, 2023. 5

[2] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 3

[3] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 3

[4] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6994–7004, 2022. 3, 7, 8

[5] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 2, 3, 4, 7, 8

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Alex Ergasti, Tomaso Fontanini, Claudio Ferrari, Massimo Bertozzi, and Andrea Prati. Mars: Paying more attention to visual attributes for text-based person search. *arXiv preprint arXiv:2407.04287*, 2024. 1, 3

[9] Shuting He, Hao Luo, Wei Jiang, Xudong Jiang, and Henghui Ding. Vgsg: Vision-guided semantic-group network for text-based person search. *IEEE Transactions on Image Processing*, 33:163–176, 2023. 3

[10] Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Attribute-guided pedestrian retrieval: Bridging person re-id with internal attribute variability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17689–17699, 2024. 1, 3

[11] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, pages 169–185. Springer, 2024. 3

[12] Boseung Jeong, Jicheol Park, and Suha Kwak. Asmr: Learning attribute-based person search with adaptive semantic margin regularizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12016–12025, 2021. 1, 3

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 3, 7, 8

[14] Jian Jia, Houjing Huang, Xiaotang Chen, and Kaiqi Huang. Rethinking of pedestrian attribute recognition: A reliable evaluation under zero-shot pedestrian identity setting. *arXiv preprint arXiv:2107.03576*, 2021. 4

[15] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023. 3

[16] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2023. 3, 7

[17] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115, 2015. 2, 3, 7

[18] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2019. 2, 3, 4, 7

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3, 7, 8

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979, 2017. 3

[21] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. *arXiv preprint arXiv:2211.13977*, 2022. 3

[22] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks, 2023. 3, 4

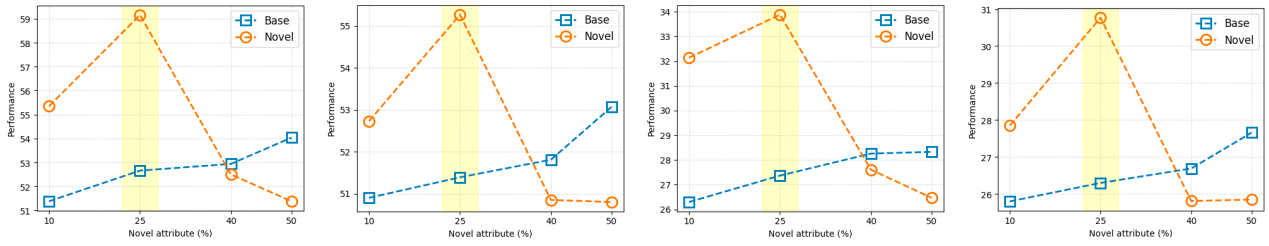
[23] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017. 2, 3, 4, 7

[24] Minjeong Park, Hongbeen Park, and Jinkyu Kim. Vita-par: Visual and textual attribute alignment with attribute prompting for pedestrian attribute recognition. 2025. 3, 5, 7, 8

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,

Open-Attribute Recognition for Person Retrieval: Finding People Through Distinctive and Novel Attributes

Supplementary Material



(a) P@1-lbl (b) P@5-lbl (c) P@1-ins (d) P@5-ins

Figure S1. Evaluation under different novel attribute ratio.

1. Detailed Explanation of OAPR

1.1. Differences Between ABPR&TBPR and OAPR

This section clarifies the fundamental differences between Attribute-based Person Retrieval(ABPR) and Text-Based Person Retrieval (TBPR) and our Open-Attribute Person Retrieval (OAPR) setting and explains why we do not include ABPR and TBPR as a baseline for our task. Although ABPR and TBPR also use text inputs, they are designed to retrieve a single target identity. Specifically, they assume that the query describes a unique person, the gallery contains multiple views of that identity, and the goal is to recover that identity. In contrast, OAPR is designed to retrieve any image that satisfies a specified set of semantic attributes. The query describes attribute combinations rather than an individual, and multiple gallery images may be valid. The goal is to retrieve all matching instances. Direct comparison between TBPR and OAPR would therefore be invalid and would misrepresent the intended behavior of both types of methods.

1.2. Difference Between POAR and OAPR

While both POAR [9] and our OAPR setting aim to handle open-vocabulary pedestrian attributes, they tackle different tasks in terms of problem formulation, supervision, and evaluation. POAR formulates the problem as an image-to-text search task. Given a pedestrian image, the goal is to generate or retrieve a textual description that lists the attributes present in the image. The main focus of POAR is to assess how well a model can produce accurate attribute text for a given image under open-vocabulary conditions and domain shifts. In contrast, OAPR is formulated as a text-to-image retrieval task. Given a set of attribute cues describing a target person, the model is required to retrieve all images of individuals matching those attributes from a gallery, regardless of whether the attributes were seen during training.

2. Open-Attribute Recognition

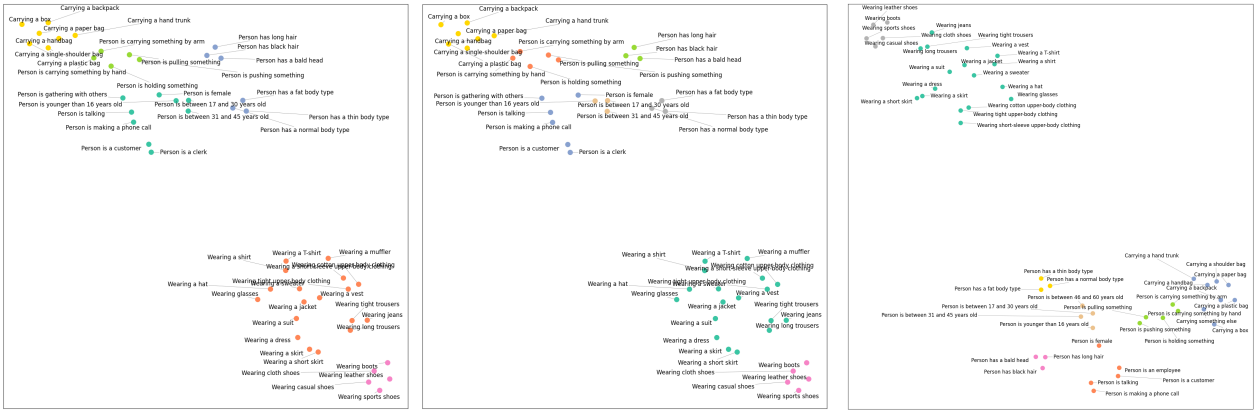
2.1. Open-Attribute Dataset

Rationale for using PAR Datasets. There exist many attribute-related datasets for person analysis; however, pedestrian attribute recognition (PAR) datasets provide the most diverse attribute categories, making them a particularly valuable source for knowledge transfer.

Analysis on the proprtion of novel attributes. We conducted a detailed analysis by varying the proportion of novel attributes per cluster (10%, 25%, 40%, and 50%) as shown in Fig. S1 on PETA. When only 10% of the attributes are designated as novel, 90% remain in the base set. Since the base split contains a large number of semantically similar attributes, it requires fine-grained discrimination among base classes and ultimately leads to lower base performance. At higher novel ratios (40-50%), the number of base attributes is substantially reduced, so the base task becomes inherently simpler. The model needs to distinguish among

Sentence-Bert MPNet E5mistral-7b

Figure S2. Ablation on Attribute Clustering Encoders.



PETA RAPv1 RAPv2

Figure S3. Visualization of Attribute Clusters Across Datasets.

fewer attribute classes and a narrower range of appearance patterns. As a result, base performance increases, reflecting reduced task difficulty rather than a fundamentally better learned representation. In contrast, the 25% novel configuration strikes a more desirable balance. The base set remains sufficiently diverse to support stable representation learning, while the novel set is large enough to induce a meaningful distribution shift, yielding a more informative and reliable evaluation of open-attribute generalization. Therefore, we adopt the 25% novel ratio as the default setting for our benchmark.

Details of Attribute Verbalization. As mentioned, each attribute is converted into a natural language description. In this section, we describe the process of transforming attribute phrases into descriptive text using GPT-3 [1], which provides richer and more diverse linguistic expressions than simple template-based methods, thereby improving attribute-text alignment. As shown in Tab. S1, we modify each attribute to be more compatible with Sentence-BERT, enhancing its ability to capture semantic similarity between attribute descriptions.

Text Encoder Analysis for Attribute Clustering We analyze the impact of the text encoder used for attribute clustering as shown in Fig. S3. Specifically, we compare three representative sentence embedding models: Sentence-BERT [6], MPNet [7], and E5mistral-7B [8] on PA-100K [5]. Sentence-BERT produces clusters that better align with human perception of attribute semantics. For instance, attributes such as "carrying backpack," "carrying handbag," "carrying shoulder bag," and "holding something in front" are grouped into a single "carrying/holding object" category by Sentence-BERT, whereas MPNet places them into separate clusters. Since these attributes all describe a person holding or carrying an object in front of the body, treating them as a coherent group is more consistent with how operators would issue attribute-based queries in practice. A similar trend is observed for head-related accessories. In Sentence-BERT, attributes like "wearing hat" and

Attribute	Attribute
Wearing a hat	Wearing a long coat
Wearing glasses	Wearing trousers
Wearing a short-sleeve shirt	Wearing shorts
Wearing a long-sleeve shirt	Wearing a skirt or dress
Wearing upper-body stride clothing	Wearing boots
Upper body clothing has a logo	Carrying a handbag
Wearing plaid upper-body clothing	Carrying a shoulder bag
Wearing upper-body splice clothing	Carrying a backpack
Wearing lower-body striped clothing	Holding objects in front
Wearing lower-body patterned clothing	Person is older than 60 years
Person is between 18 and 60 years old	Person is younger than 18 years old
Person is female	Facing front
Facing side	Facing back

(a) PA-100K [5]

Attribute	Attribute
Wearing a hat	Wearing trousers
Wearing a muffler	Wearing leather shoes
Wearing sunglasses	Wearing sandals
Has long hair	Wearing shoes
Wearing casual upper body clothing	Wearing sneakers
Wearing formal upper body clothing	Carrying a backpack
Wearing a jacket	Carrying something
Upper body clothing has a logo	Carrying a messenger bag
Wearing a plaid upper body clothing	Carrying plastic bags
Wearing a short-sleeve upper body clothing	Person is less than 30 years old
Upper body clothing has thin stripes	Person is less than 45 years old
Wearing a t-shirt	Person is less than 60 years old
Wearing other upper body clothing	Person is older than 60
Wearing a V-neck upper body clothing	Person is male
Wearing casual lower body clothing	
Wearing formal lower body clothing	
Wearing jeans	
Wearing shorts	
Wearing a short skirt	

(b) PETA [2]

Attribute	Attribute
Person has a bald head	Wearing leather shoes
Person has long hair	Wearing sports shoes
Person has black hair	Wearing boots
Wearing a hat	Wearing cloth shoes
Wearing glasses	Wearing casual shoes
Wearing a muffler	Carrying a backpack
Wearing a shirt	Carrying a single-shoulder bag
Wearing a sweater	Carrying a handbag
Wearing a vest	Carrying a box
Wearing a T-shirt	Carrying a plastic bag
Wearing cotton upper-body clothing	Carrying a paper bag
Wearing a jacket	Carrying a hand trunk
Wearing a suit	Person is younger than 16 years old
Wearing tight upper-body clothing	Person is between 17 and 30 years old
Wearing a short-sleeve upper-body clothing	Person is between 31 and 45 years old
Wearing long trousers	Person is female
Wearing a skirt	Person has a fat body type
Wearing a short skirt	Person has a normal body type
Wearing a dress	Person has a thin body type
Wearing jeans	Person is a customer
Wearing tight trousers	Person is a clerk
Person is making a phone call	Person is talking
Person is gathering with others	Person is holding something
Person is pushing something	Person is pulling something
Person is carrying something by arm	Person is carrying something by hand

(c) RAPv1 [3]

Attribute	Attribute
Person has a bald head	Wearing leather shoes
Person has long hair	Wearing sports shoes
Person has black hair	Wearing boots
Wearing a hat	Wearing cloth shoes
Wearing glasses	Wearing casual shoes
Wearing a shirt	Carrying a backpack
Wearing a sweater	Carrying a shoulder bag
Wearing a vest	Carrying a handbag
Wearing a T-shirt	Carrying a box
Wearing cotton upper-body clothing	Carrying a plastic bag
Wearing a jacket	Carrying a paper bag
Wearing a suit	Carrying a hand trunk
Wearing tight upper-body clothing	Carrying something else
Wearing short-sleeve upper-body clothing	Person is younger than 16 years old
Wearing long trousers	Person is between 17 and 30 years old
Wearing a skirt	Person is between 31 and 45 years old
Wearing a short skirt	Person is between 46 and 60 years old
Wearing a dress	Person is female
Wearing jeans	Person has a fat body type
Wearing tight trousers	Person has a normal body type
Person has a thin body type	Person is a customer
Person is an employee	Person is making a phone call
Person is talking	Person is holding something
Person is pushing something	Person is pulling something
Person is carrying something by arm	Person is carrying something by hand

(d) RAPv2 [4]

Table S1. Verbalized attribute phrases from four PAR datasets used in our experiments.

"wearing glasses" are assigned to the same cluster, reflecting that they are both accessories worn on the head or face. In contrast, E5mistral-7B, despite its much larger capacity, often groups "wearing hat" and "wearing glasses" together with lower-body attributes such as "boots," "trousers," "shorts," and "skirt/dress." This indicates that E5mistral-7B tends to over-cluster heterogeneous attributes that merely co-occur in generic text, leading to clusters that are less structured and less interpretable for pedestrian retrieval. Since OAPR relies on such clusters to define meaningful base/novel partitions, we adopt Sentence-BERT as the default encoder for constructing the Open-Attribute Dataset.

Analysis on Attribute Clusters. We further visualize the attribute clusters produced by Sentence-BERT across the rest of three datasets (e.g. PETA [2], RAPv1 [3] and RAPv2 [4]). Despite the datasets differing substantially in scale, annotation style, and attribute vocabulary, Sentence-BERT forms consistent and semantically coherent structures. These patterns indicate that the attribute embedding space learned by Sentence-BERT remains stable across datasets, producing human-aligned clusters that are suitable for defining base–novel partitions. The consistency across datasets further supports the use of Sentence-BERT as the default encoder for constructing the Open-Attribute Dataset.

Base and Novel Attribute Splits. We present the base and novel attribute divisions for PA-100K [5], PETA [2], RAPv1 [3], and RAPv2 [4]. As shown in Tab. S2, we partition each cluster into base and novel classes to ensure that our model can effectively transfer from base attributes to novel attributes. It allows us to evaluate zero-shot generalization to unseen attribute names that lie near the training distribution, under a controlled setting where (i) the base and novel sets are disjoint in terms of labels, but (ii) both

Cluster No.	Base		Novel
0	Backpack HandBag	HoldObjectsInFront	ShoulderBag
1	Trousers boots	Shorts	Skirt&Dress
2	Age18-60 Female	AgeLess18	AgeOver60
3	Front	Side	Back
4	UpperPlaid LowerStripe	UpperSplice LowerPattern	UpperStride LongCoat
5	LongSleeve	UpperLogo	ShortSleeve
6	Glasses		Hat

(a) PA-100K [5]

Cluster No.	Base		Novel
0	accessoryHat hairLong	accessoryMuffler	accessorySunglasses
1	upperBodyCasual upperBodyPlaid upperBodyThinStripes upperBodyOther	upperBodyFormal upperBodyShortSleeve upperBodyTshirt upperBodyJacket	upperBodyLogo upperBodyVNeck lowerBodyFormal
2	lowerBodyCasual lowerBodyJeans	lowerBodyShorts	lowerBodyShortSkirt lowerBodyTrousers
3	personalLess45 personalLarger60	personalLess60	personalLess30 personalMale
4	footwearLeatherShoes footwearSneaker	footwearSandals	footwearShoes
5	carryingBackpack carryingOther	carryingPlasticBags	carryingMessengerBag

(b) PETA [2]

Cluster No.	Base		Novel
0	ub-Shirt ub-TShirt ub-SuitUp lb-Jeans lb-Skirt	ub-Vest ub-Jacket ub-Tight lb-TightTrousers lb-Dress	ub-Sweater ub-Cotton ub-ShortSleeve lb-LongTrousers lb-ShortSkirt
1	action-Talking action-Pusing action-CarrybyArm	action-Gathering action-Pulling	action-Calling action-CarrybyHand action-Holding
2	Female	Customer	Clerk
3	shoes-Leather shoes-Boots	shoes-Sport shoes-Casual	shoes-Cloth
4	hs-BaldHead hs-Hat	hs-BlackHair hs-Glasses	hs-LongHair hs-Muffler
5	attach-SingleShoulderBag attach-PlasticBag attach-HandTrunk	attach-HandBag attach-PaperBag	attach-Backpack attach-Box
6	AgeLess16	Age31-45	Age17-30
7	BodyFat	BodyNormal	BodyThin

(c) RAPv1 [3]

Cluster No.	Base		Novel
0	hs-Hat ub-Shirt ub-TShirt ub-Jacket ub-Tight lb-ShortSkirt	hs-Glasses ub-Vest ub-Cotton ub-SuitUp lb-LongTrousers lb-Skirt	lb-TightTrousers ub-Sweater ub-ShortSleeve lb-Dress lb-Jeans
1	BodyNormal	BodyThin	BodyFat
2	attachment-Backpack attachment-HandBag attachment-PaperBag	attachment-ShoulderBag attachment-PlasticBag	attachment-Box attachment-HandTrunk
3	hs-BlackHair	hs-LongHair	hs-BaldHead
4	action-Pushing action-CarryingByHand	action-CarryingByArm	action-Holding action-Pulling
5	action-Calling action-Gathering	action-Talking	Customer Employee
6	AgeLess16 Age31-45	Age17-30 Age46-60	Female
7	shoes-Leather shoes-Casual	shoes-Boots	shoes-Sports shoes-Cloth

(d) RAPv2 [4]

Table S2. The base and novel attributes split on Open-Attribute Dataset.

sets cover comparable semantic categories (clothing, accessories, age, pose, etc.).

In the main paper, we additionally report qualification results on external novel attributes that are not part of any PAR annotation set. While Tab. S2 represents base and novel attributes obtained by splitting and clustering the original PAR attributes, this setting is still confined to the PAR attribute vocabulary. To examine how our model behaves on attributes that go beyond this vocabulary, we further evaluate retrieval performance on completely new attribute queries such as "bicycle," "helmet," and "stroller." These attributes are not annotated in PA-100K, PETA, RAPv1, or RAPv2 and are never used as training labels, but they form meaningful attribute-like queries for pedestrian images (e.g., "person with a helmet," "person pushing a stroller").

2.2. Evaluation Metrics

Let N be the number of gallery images and M the number of attributes. We denote the binary attribute label matrix as

$$\mathbf{Y} \in \{0,1\}^{N\times M}, \qquad Y_{n,m} \in \{0,1\}. \tag{1}$$

For each image–attribute pair, the model predicts a similarity score

$$\mathrm{sim}_{n,m} \in [0,1], \tag{2}$$

which we interpret as the probability that image n possesses attribute m. For a fixed attribute-set size k, we consider all possible combinations of k attributes and their binary states (present or absent). Let

$$\mathcal{C}_k = \{(S, \mathbf{q})\} \tag{3}$$

denote the set of all such queries, where $S \subset \{1,\dots,M\}$, $|S| = k$ is the selected attribute index set, $\mathbf{q} = (q_j)_{j\in S} \in \{0,1\}^k$ specifies whether each attribute $j \in S$ must be present ($q_j = 1$) or absent ($q_j = 0$). For a given query

Figure S4. Additional visualization of person retrieval according to the given attributes.

$(S, \mathbf{q})$, we define the likelihood that image x_n satisfies all required attribute conditions as

$$P_{S,\mathbf{q}}(n) = \prod_{j \in S} \left(\mathbb{I}[q_j = 1] \cdot \mathrm{sim}_{n,j} + \mathbb{I}[q_j = 0] \cdot (1 - \mathrm{sim}_{n,j}) \right) \tag{4}$$

, where $n = 1, \ldots, N$. We then retrieve the top-R images with the highest query likelihood:

$$\mathrm{TopR}(S, \mathbf{q}) = \underset{n \in \{1,\ldots,N\}}{\arg \mathrm{topR}} \; P_{S,\mathbf{q}}(n). \tag{5}$$

Precision@K for label (P@K-lbl). For a query $(S, \mathbf{q})$, the label-level precision evaluates attribute correctness averaged over the k attributes and the R retrieved images:

$$\mathrm{Prec}^{\mathrm{lbl}}_{k,R}(S, \mathbf{q}) = \frac{1}{Rk} \sum_{n \in \mathrm{TopR}(S,\mathbf{q})} \sum_{j \in S} \mathbb{I}\big[Y_{n,j} = q_j\big]. \tag{6}$$

The final label-level precision is the average over all query combinations:

$$\mathrm{Prec}^{\mathrm{lbl}}_{k,R} = \frac{1}{|\mathcal{C}_k|} \sum_{(S,\mathbf{q}) \in \mathcal{C}_k} \mathrm{Prec}^{\mathrm{lbl}}_{k,R}(S, \mathbf{q}). \tag{7}$$

Precision@K for instance (P@K-ins). Instance-level precision evaluates how often the retrieved images fully satisfy the attribute set:

$$\mathrm{Prec}^{\mathrm{ins}}_{k,R}(S, \mathbf{q}) = \frac{1}{R} \sum_{n \in \mathrm{TopR}(S,\mathbf{q})} \mathbb{I}\big[\forall j \in S,\; Y_{n,j} = q_j\big]. \tag{8}$$

Averaging over all combinations yields:

$$\mathrm{Prec}^{\mathrm{ins}}_{k,R} = \frac{1}{|\mathcal{C}_k|} \sum_{(S,\mathbf{q}) \in \mathcal{C}_k} \mathrm{Prec}^{\mathrm{ins}}_{k,R}(S, \mathbf{q}). \tag{9}$$

3. Experiments

Body and Background Related Classes. We set the background related class as follows: 'airplane', 'bed', 'bedclothes', 'bench', 'bicycle', 'bird', 'boat', 'book', 'bottle', 'building', 'bus', 'cabinet', 'car', 'cat', 'ceiling', 'chair', 'computer', 'cow', 'cup', 'curtain', 'dog', 'door', 'fence', 'floor', 'flower', 'food', 'grass', 'ground', 'horse', 'keyboard', 'light', 'motorbike', 'mountain', 'mouse', 'plate', 'platform', 'potted plant', 'road', 'rock', 'sheep', 'shelves', 'sidewalk', 'sign', 'sky', 'snow', 'sofa', 'table', 'track', 'train', 'tree', 'truck', 'tv monitor', 'wall', 'water', 'window', 'wood'.

More visualization results. We present additional qualitative examples of attribute-based person retrieval under the OAPR setting. A model trained on RAPv2 is evaluated for cross-dataset retrieval using samples from PETA. As shown Fig. S4, the qualitative results demonstrate that it can accurately retrieve individuals based on a wide range of attribute categories, including those that were absent during training.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[2] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 789–792, 2014. 3, 4

[3] Dangwei Li, Xiaotang Chen, and Kaiqi Huang. Multi-attribute learning for pedestrian attribute recognition in surveillance scenarios. In *ACPR*, pages 111–115, 2015. 3, 4

[4] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE transactions on image processing*, 28(4):1575–1590, 2019. 3, 4

[5] Xihui Liu, Haiyu Zhao, Maoqing Tian, Lu Sheng, Jing Shao, Junjie Yan, and Xiaogang Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2017. 2, 3, 4

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 2

[7] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020. 2

[8] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, 2024. 2

[9] Yue Zhang, Suchen Wang, Shichao Kan, Zhenyu Weng, Yigang Cen, and Yappeng Tan. Poar: Towards open vocabulary pedestrian attribute recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, 2023. 1

Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 6, 7, 8

[28] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019. 4

[29] Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. Pedestrian attribute recognition via clip-based prompt vision-language fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 3, 5, 7, 8

[30] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5314–5322, 2022. 3

[31] Peng Wu, Xuerong Zhou, Guansong Pang, Yujia Sun, Jing Liu, Peng Wang, and Yanning Zhang. Open-vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18297–18307, 2024. 3

[32] Xi Yang, Xiaoqi Wang, and Dong Yang. Improving cross-modal constraints: Text attribute person search with graph attention networks. *IEEE Transactions on Multimedia*, 26: 2493–2503, 2024. 1, 3

[33] Yajing Zhai, Yawen Zeng, Zhiyong Huang, Zheng Qin, Xin Jin, and Da Cao. Multi-prompts learning with cross-modal alignment for attribute-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6979–6987, 2024. 1

[34] Yue Zhang, Suchen Wang, Shichao Kan, Zhenyu Weng, Yigang Cen, and Yappeng Tan. Poar: Towards open vocabulary pedestrian attribute recognition. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, 2023. 3, 5, 7, 8

[35] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 7

[36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 3, 5, 7

[37] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2023. 4

[38] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 3

[39] Jun Zhu, Jiandong Jin, Zihan Yang, Xiaohao Wu, and Xiao Wang. Learning clip guided visual-text fusion transformer for video-based pedestrian attribute recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2626–2629, 2023. 3

[40] Jianqing Zhu, Liu Liu, Yibing Zhan, Xiaobin Zhu, Huanqiang Zeng, and Dacheng Tao. Attribute-image person re-identification via modal-consistent metric learning. *International Journal of Computer Vision*, 131(11):2959–2976, 2023. 1