Attribute-Guided Pedestrian Retrieval: Bridging Person Re-ID with Internal Attribute Variability

Yan Huang[1] Zhang Zhang[1,2*] Qiang Wu[3] Yi Zhong[4] Liang Wang[1,2*]
[1] Center for Research on Intelligent Perception and Computing (CRIPAC),
State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS),
Institute of Automation, Chinese Academy of Sciences, Beijing China.
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS).
[3] School of Electrical and Data Engineering, University of Technology Sydney, Australia.
[4] School of Information and Electronics, Beijing Institute of Technology, China.
huangyan.750@outlook.com, {zzhang,wangliang}@nlpr.ia.ac.cn,
qiang.wu@uts.edu.au, yi.zhong@bit.edu.cn

Abstract

In various domains such as surveillance and smart retail, pedestrian retrieval, centering on person re-identification (Re-ID), plays a pivotal role. Existing Re-ID methodologies often overlook subtle internal attribute variations, which are crucial for accurately identifying individuals with changing appearances. In response, our paper introduces the Attribute-Guided Pedestrian Retrieval (AGPR) task, focusing on integrating specified attributes with query images to refine retrieval results. Although there has been progress in attribute-driven image retrieval, there remains a notable gap in effectively blending robust Re-ID models with intra-class attribute variations. To bridge this gap, we present the Attribute-Guided Transformer-based Pedestrian Retrieval (ATPR) framework. ATPR adeptly merges global ID recognition with local attribute learning, ensuring a cohesive linkage between the two. Furthermore, to effectively handle the complexity of attribute interconnectivity, ATPR organizes attributes into distinct groups and applies both inter-group correlation and intra-group decorrelation regularizations. Our extensive experiments on a newly established benchmark using the RAP dataset [32] demonstrate the effectiveness of ATPR within the AGPR paradigm.

1. Introduction

Pedestrian retrieval plays a crucial role in various domains, including security and smart retail. At its core lies the technique of person re-identification (Re-ID), which focuses on identifying individuals across camera views [19, 23, 63, 69–71]. Although existing Re-ID methods demonstrate notable accuracy and scalability [16, 17, 22, 40, 55, 63], they tend to overlook subtle, individual attribute variations, which are essential for precise identification.

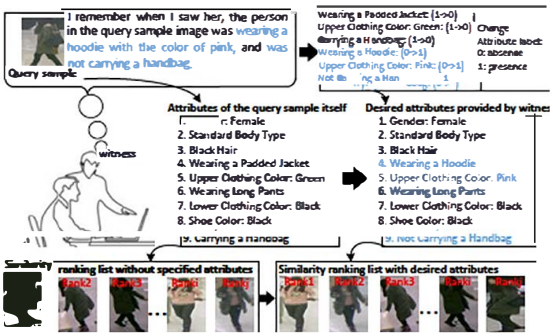

Figure 1. An illustration of the AGPR task. The figure showcases how the additional attributes remembered by a witness but not consistent with the queried image. Initial similarity rankings without specific attributes are contrasted with attribute-guided rankings based on the witness's attribute descriptions.

Consider a series of subway thefts as an example. The police presents witnesses with a mugshot of the suspect. Due to the thief's potential disguises or the limited field of view (FoV) of the mugshot, witnesses might offer varied attribute descriptions. For instance, one might mention a "green scarf and a tattoo on his left hand," while another recalls "sunglasses and a blue jacket." This necessitates the development of an efficient method capable of merging the old mugshot with attributes provided by witnesses to identify the suspect in CCTV footage. This retrieval with mul-

*Corresponding authors: Liang Wang, Zhang Zhang

timodal queries is no doubt an effective method to improve the searching results, underscoring the importance of integrating Re-ID systems with detailed attribute information. We term this innovative task as the Attribute-Guided Pedestrian Retrieval (AGPR) task. Fig. 1 illustrates an example.

Attribute-driven image retrieval has advanced, particularly in fashion [39, 53] and face retrieval [6, 12, 38, 65]. However, these innovations often do not address intra-class attribute variability and lack focus on consistent ID recognition. Similarly, while text-based person searches [1, 7, 48, 50] bridge textual descriptors and visual content, they fall short in handling the complexities of attribute variations within an ID. While some studies have investigated enhancing Re-ID systems by integrating attributes during the training phase [36, 42, 52], these methods generally utilize attribute labels to boost performance rather than addressing intra-ID attribute variations. Consequently, these approaches are typically confined to contexts where intra-class attributes do not vary.

Confronted with these issues, a clear gap becomes evident: the need to develop a paradigm that seamlessly integrates person Re-ID with the variability of internal attributes. We emphasize that while the cornerstone of person Re-ID remains the identification of the correct ID, integrating attribute guidance can enhance the specificity of retrieval results. For instance, searching for an individual 'wearing a red hat' should yield not just any individual, but specifically the person depicted in the queried image.

The primary challenge in AGPR lies in integrating nuanced internal attribute variations into person Re-ID matching models. In real-world scenarios, individuals may change their appearance, which can include alterations in clothing, hairstyle, or accessories. Such variability poses a significant challenge, as these changes can lead to intra-ID confusion, where the same individual is misidentified across different camera perspectives or over time. Additionally, ensuring the Re-ID system's robustness against attribute variability, while maintaining high accuracy and efficiency, is another critical hurdle. Moreover, over-reliance on specific attributes combinations could diminish the system's effectiveness. In diverse environments, individuals may present with a wide range of unseen attribute combinations, further compound the complexity of the task.

Current methodologies face significant limitations when addressing the above challenges. For instance, existing clothing-change Re-ID models primarily concentrate on adapting to variations in apparel [10, 18, 24, 28, 37, 60, 61]. Yet, the challenge of establishing the ID of a target individual without constraining specific attribute changes remains formidable. Moreover, while existing attribute recognition frameworks are adept at identifying specific attributes [49, 51, 54], they typically lack a comprehensive approach to learning and integrating ID information. These challenges highlight a crucial gap in existing pedestrian retrieval systems – the need for a more holistic approach that considers the capability to pinpoint targets with specified attributes.

To address these challenges, this paper introduces the Attribute-Guided Transformer-based Pedestrian Retrieval (ATPR) framework that effectively integrates global ID recognition and local attribute learning. This dual-faceted strategy ensures that while the system recognizes individual attribute changes, it still maintains a strong foundation in ID recognition. ATPR is particularly adept at handling intra-ID confusion by creating a coherent connection between a person's attributes and their ID, even amidst alterations. Additionally, the framework employs attribute correlation/decorrelation regularization techniques to manage the intricate relationships between different attributes. This regularization helps in distinguishing individuals based on their unique attribute features. Furthermore, the ATPR framework's efficacy in handling intra-ID attribute variability is validated by extensive experiments on a new evaluation benchmark using the RAP dataset [32]. Our approach allows for the specification of attributes during the retrieval process, enabling more precise and targeted Re-ID. The primary contributions of this paper are outlined as follows:

- We introduce the novel AGPR task in pedestrian retrieval, utilizing fine-grained attribute for more targeted image retrieval. This approach enhances search precision in applications requiring specific attribute identification.
- We present the ATPR framework that effectively combines global ID recognition with local attribute learning, addressing intra-class attribute variation. It includes inter-group attribute correlation and intra-group decorrelation regularizations to manage attribute interconnectedness.
- A new AGPR evaluation benchmark is established using the RAP dataset [32], featuring rich attribute annotations and significant variability. Extensive experiments demonstrate our ATPR's ability to enhance pedestrian image retrieval by focusing on specific attribute variations.

2. Related work

Attribute-Guided Image Retrieval is a key area in vision research [2, 26, 44, 64, 74], with notable advancements in fashion image retrieval and face image retrieval. Within *Fashion Image Retrieval*, Vo et al. [53] innovated a system where queries consist of an input image paired with textual descriptions detailing desired modifications. Liu et al. [39] capitalized on pre-existing vision-and-language frameworks to adapt visual features based on natural language inputs. Both works utilize a shared architecture to encode the query image, textual conditions, and target image. Other notable contributions, such as [8, 11, 29, 30, 68], empowered users or embedded natural language systems to offer iterative feedback on retrieved items, thus refining the search outcomes. Turning to *Face Image Retrieval*, Gupta

et al. [12] harnessed user feedback to categorize images as either aligning with or deviating from their mental picture, subsequently guiding the face images during retrieval. Zaeemzadeh *et al.* [65] enhanced the input face query utilizing facial attributes. Liu *et al.* [38] adopted facial attributes to steer the generation of face visuals.

Contrary to the above works, which often overlooks the need for ID consistency or attribute correlations, our AGPR places emphasis on the detailed integration of individual identity recognition with correlated person attributes. Our AGPR differentiates itself in the image retrieval domain by specifically addressing real-world necessities through a thorough and precise person ID matching process.

Text-Based Person Search primarily utilizes textual descriptions to identify individuals in images and videos [1, 3, 7, 41, 48, 50, 57]. Wu *et al.* [57] enhanced the learning of fine-grained cross-modality connections through a color-reasoning sub-task. Gao *et al.* [7] synchronized person image sub-regions with their corresponding textual phrase descriptions. Jiang *et al.* [27] showcased the applicability of the full CLIP model [46] to the text-to-image person retrieval task with minimal fine-tuning. While text-based person search tackles a cross-modal challenge (translating textual attributes into visual cues), our AGPR intricately merges Re-ID with specific person attributes. Rather than depending exclusively on textual descriptions, our AGPR places more emphasis on visual cues to identify individuals, further refining retrieval based on detailed attributes.

Person Re-ID with Attribute aims to enhance identity learning through multi-task learning [36, 42, 52]. Lin *et al.* [36] learned a CNN embedding for both person Re-ID and attribute recognition via an attribute re-weighting module. Schumann *et al.* [47] incorporated the semantic details of attributes into the CNN learning process for person Re-ID. Zhang *et al.* [66] devised an attribute attentional block to harness fine-grained attribute attention, further enhancing Re-ID performance. However, these approaches overlook instances where individuals may alter an attribute.

Clothing-Change Person Re-ID considers scenarios where a person might alter their clothing [10, 13, 15, 21, 28, 37, 56, 61, 67]. These approaches predominantly focus on adapting the network to overcome dress variations, thereby ensuring persistent recognition despite clothing changes [21, 37]. Our AGPR distinguishes itself by not only identifying individuals but also enhancing the retrieval precision through an in-depth attribute-centric analysis.

Vehicle Re-ID with Attribute is centered on identifying vehicles, often using attributes like mark, model, color, or license plates [33, 62, 73]. Vehicle attributes are usually distinct and less prone to rapid change. In our AGPR, the specification and retrieval based on attributes are more challenging due to the frequent and unpredictable nature of human attribute changes.

Table 1. Attribute groupings and their respective descriptions.

Group	Description	Group	Description
a^1	Basic Personal Attributes	a^2	Head and Shoulder
a^3	Upper Body Clothing	a^4	Lower Body Clothing
a^5	Footwear	a^6	Accessory

3. Methodology

3.1. Input data

As delineated in Fig. 2, our ATPR processes pedestrian images combined with corresponding attributes. We manually categorize these attributes into six groups $\{a^i|i = 1, 2, ..., 6\}$, each containing distinct attributes a^i_j, where j represents the count of attributes in a particular group. The categorization is detailed in Tab. 1.

Before feeding pedestrian images to our model, they are pre-processed with Openpose [4] to identify anatomical keypoints, then segmented into 16×16 non-overlapping patches $\{x_p|p = 1, 2, ..., N\}$. Each patch (or token) is converted into a 1D vector and undergoes linear projection for dimensionality D. These vectors form token embeddings for the transformer layers [5], with an additional [class] token embedding to discern global features.

3.2. Attribute Embedding

Patch-Attribute Association. Prior to processing token embeddings through transformer layers, each token is associated with a relevant attribute based on keypoints identified via Openpose in the patch. For a patch with N_k keypoints ($0 \leq N_k \leq K$, where K is the maximum number of keypoints and 0 signifies no keypoints), we calculate the Euclidean distance from each keypoint to the patch's center. The closest keypoint determines the token's linked attribute group, such as a^1 for keypoints on the head or shoulder.

Attribute Embedding. We introduce Attribute Embedding (AE), a learnable 1D vector, to integrate attribute information into token embeddings. This approach is inspired by the position and camera/viewpoint embeddings as outlined in [5, 14]. Unlike these methods, our AE specifically maps tokens to relevant attributes for the AGPR task. Specifically, attributes are categorized into six groups (a^i) (see Tab. 1). Groups a^1 and a^6, representing basic personal attributes and accessories, are initialized as AE A^1 and A^6, respectively, and added to the [class] token embedding (x_{cls}):

$$\mathcal{Z}_0^0 = x_{cls} + \sum_{j=1}^{j_1} A_j^1 + \sum_{j=1}^{j_6} A_j^6, \tag{1}$$

where $A_j^i \in \mathbb{R}^D$ represents the learnable AE associated with the $j^{\text{-th}}$ attribute in the $i^{\text{-th}}$ attribute group. j_1 and j_6 denote the total counts of attributes in groups a^1 and a^6, respectively (j_6=0 if no accessory).

As illustrated in Fig. 2, other token embeddings possess corresponding local body attributes linked to groups a^i

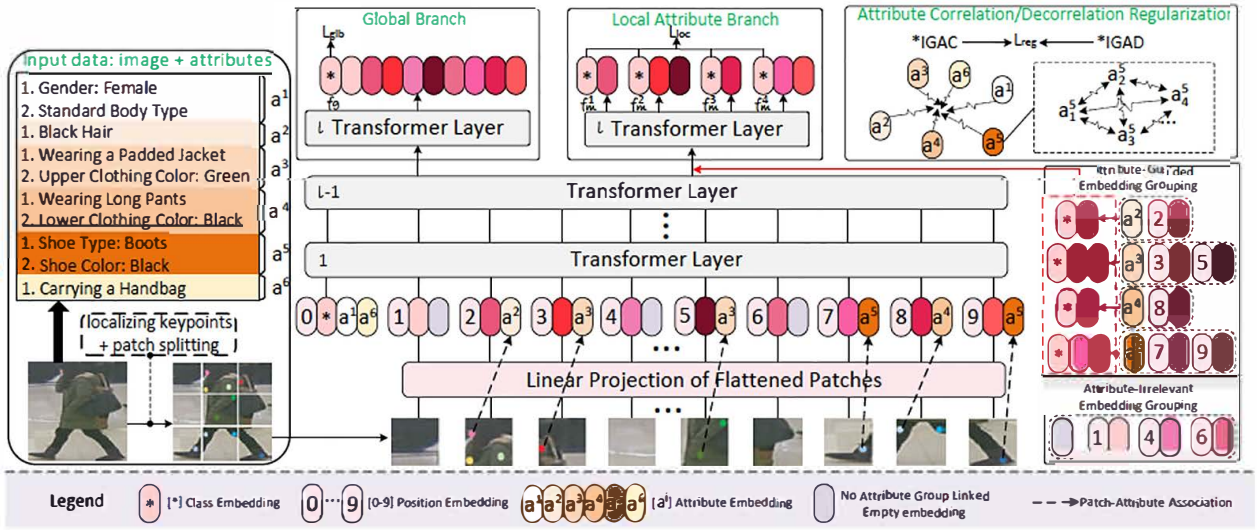

Figure 2. The AGPR architecture. The figure highlights the processing of input data, consisting of an image and its associated attributes. Key elements like global branch, local attribute branch, AEG, IGAC and IGAD, and attribute and position embeddings are illustrated.

where $i \in \{2,3,4,5\}$, or they might not have any linkage if no keypoints are detected on the patch. For these embeddings, we initialize the learnable AE A^i and add it to the p-th token embedding as:

$$\mathcal{Z}_0^p = x_p + \mathbb{I} \cdot \sum_{j=1}^{j_i} A_j^{p \to i}, i \in \{2,3,4,5\}, \quad (2)$$

where $\mathbb{I}$ is assigned the value 1 if the patch $\{x_p | p = 1,2,...,N\}$ has a corresponding linked attribute group (represented as $p \to i$ in the equation), and 0 otherwise, where N represents the total number of patches.

Lastly, the sequences provided as the input to transformer layers are expressed as:

$$\mathcal{Z}_0 = [\mathcal{Z}_0^0; \mathcal{Z}_0^1; \mathcal{Z}_0^2; ...; \mathcal{Z}_0^N] + \mathcal{P}, \quad (3)$$

where $\mathcal{P} \in \mathbb{R}^{(N+1)\times D}$ refers to the position embedding. Mirroring the behavior in ViT [5], this position embedding is also designed to be learnable.

3.3. Global and Local Attribute learning

The sequence $\mathcal{Z}_0$ undergoes l transformer layers to derive feature representations. Inspired by [14], we implement two branches atop the output of the l-1-th transformer layer.

Global Branch. As shown in Fig. 2, the [class] token output acts as a global feature representation, denoted as f_g. Aligning with the conventional ViT-based Re-ID methodology [14], both the triplet loss (L_{trip}) and ID loss (*i.e.*, cross-entropy loss, L_{id}) are applied on global features.

$$L_{glb} = L_{id}(f_g) + L_{trip}(f_g); \quad (4)$$

Local Attribute Branch. Given the intrinsic attribute variation in our AGPR task, we craft a local attribute branch adept at intelligently assimilating local attribute information alongside ID recognition. This design ensures that the network effectively adapts to changes in an individual's attributes, maintaining the crucial link between evolving attributes and the core ID information, thereby fulfilling the objectives of the AGPR task. Specifically, assuming the hidden feature of the l-1-th transformer layer is denoted as $\mathcal{Z}_{l-1} = [\mathcal{Z}_{l-1}^0; \mathcal{Z}_{l-1}^1; \mathcal{Z}_{l-1}^2; ...; \mathcal{Z}_{l-1}^N]$, our Attribute-guided Embedding Grouping (AEG) amalgamates these hidden features into groups, based on their corresponding attribute embeddings. For instance, as indicated in Fig. 2, a^3 possesses two correlated hidden features merged together. Tokens paired with empty embeddings are disregarded in the local attribute branch. The token $\mathcal{Z}_{l-1}^0$ (i.e., output of the [class] token post the l-1 transformer layer) is appended to each group and subsequently processed by the l-th Transformer layer of the local attribute branch. Notably, since a^1 and a^6 represent global attributes merged with the [class] token, they are excluded from the local attribute branch.

For the local attribute features (f_m), we deploy a binary cross-entropy loss L_{bin} which predicts whether the attribute should be marked as 1 or 0, based on its presence or absence in each group of the local branch. Moreover, ID and triplet loss further bolster the learning of distinct ID features within local attribute zones:

$$L_{loc} = \frac{1}{M} \sum_{i=1}^{M} (L_{id}(f_m^i) + L_{trip}(f_m^i) + L_{bin}(f_m^i)), \quad (5)$$

Functionality of the local attribute branch. The local attribute branch operates on a dual learning mechanism, simultaneously acquiring knowledge about specific attributes and ID information. This design ensures the network is not solely reliant on static attributes. When attributes undergo

changes, the foundational knowledge of the ID, reinforced by integrating the [class] token's output into every attribute group post-AEG, aids in achieving consistent retrieval.

3.4. Attribute Regularization

Correlations within attribute groups often play a critical role in vision tasks [25, 59]. For inter-group correlations, such connections provide invaluable contextual insights. For instance, awareness of other correlated attributes, such as pants type, can augment the accuracy in identifying a specific footwear type. Conversely, intra-group attributes do not necessarily exhibit strong correlations always. For example, within the attribute group a^3, presented in Fig. 2 which illustrates upper body attributes, a "green" attribute does not automatically imply its association with a "jacket". Consequently, dissociating these intra-group attributes is imperative. To navigate these complexities, we devise a novel regularization mechanism combining both Inter-Group Attribute Correlation (IGAC) and Intra-Group Attribute Decorrelation (IGAD). This dual tactic seeks to optimize the balance between leveraging inter-group correlations for enhanced recognition accuracy while ensuring intra-group attributes remain uncorrelated, structured as:

$$L_{reg} = \sum_{i=1}^{6} \sum_{\substack{q,k \in A^i \\ q \neq k}} \left\| A_k^i - A_q^i \right\|_2 - \lambda \sum_{\substack{i,j=1 \\ i \neq j}}^{6} \| \frac{1}{|A^i|} \sum_{k \in A^i} A_k^i - \frac{1}{|A^j|} \sum_{q \in A^j} A_q^j \|_2. \tag{6}$$

The strategy encompasses two main objectives. First, within each attribute group (e.g., A^i), we compute pairwise L2 distances among all unique attributes to ensure that intra-group attributes remain decorrelated. This ensures that attributes within the same group are less interconnected. Second, when considering different attribute groups, we determine the average representation for each group and compute the L2 distances between these attribute prototypes. Using a negative sign, this term aspires to minimize discrepancies between distinct attribute groups, accentuating their correlations. The balancing factor, λ, mediates the dual goals of intra-group decorrelation and inter-group correlation.

Introducing IGAC and IGAD regularizations offers a more nuanced control and helps the network understand the relationship between attributes. This is pivotal because even if one attribute changes, the network can assess its relationship with other attributes and make informed decisions about the ID association.

3.5. Total Objective

Total Objective. The final loss function is:

$$L = L_{glb} + \alpha L_{loc} + \beta L_{reg}, \tag{7}$$

where α and β serve as hyperparameters determining the weights of individual losses.

During inference, when provided with a query image accompanied by specific attributes (potentially differing from the attributes evident in the image), we concatenate the global and the mean local attribute features $f = [f_g, \overline{\sum_{i=1}^{M} f_m^i}]$ to form the final representation. As delineated in Sec. 3.3, our local attribute branch can associate the newly introduced attributes with the original ID, effectively accomplishing the AGPR task.

4. Experiments

4.1. New AGPR benchmark based on RAP

Our experimental dataset for AGPR research meets three key requirements: suitability for Re-ID tasks, intra-class attribute variation, and a rich set of attribute labels. The RAP dataset [32] is uniquely suited for this purpose, offering 2,589 pedestrian IDs across 26,638 images. The training set includes 13,178 images with 1,295 unique IDs, and the test set comprises 13,460 images with another 1,294 unique IDs, including 7,202 query images and 6,258 gallery images. Each image is annotated with 156 attribute labels, from which we selected 86 biological and appearance attributes, categorized into six groups as outlined in Tab. 1. These attributes, represented in an 86-dimensional vector (0/1 indicating the absence/presence of an attribute), cover aspects like gender, age, body type, clothing specifics, and accessories. Analysis of the RAP dataset shows significant intra-ID attribute variation, with each ID averaging at least four attribute changes, essential for our AGPR task. Note that an arXiv paper [15] annotates a clothing-change Re-ID dataset with 20 pedestrian attributes per image, but these labels are not public, and the attribute set seems limited.

4.2. Implementation Details

All person images are resized to 256×128 pixels. Following the established ViT-based Re-ID baseline protocols [14], the training images undergo augmentation that includes random horizontal flipping, padding, random cropping, and random erasing [72]. A patch size of 16×16 with a stride of 16 is used, resulting in 128 patches for each image as input. The batch size is set to 64, comprising 4 images per ID. The SGD optimizer is employed with a momentum of 0.9 and weight decay of 1e-4. The learning rate is initialized at 0.008 and follows a cosine decay schedule. Initial ViT weights are pre-trained on ImageNet-21K and then fine-tuned on ImageNet-1K. In Eq. 6, λ is set to 1. In Eq. 7, α is set to 1 and β is set to 0.05.

Evaluation Protocol. We employ standard cmc (e.g., rank-1) and mAP metrics for evaluation. Unlike traditional settings that rely solely on the ID label for matching, the AGPR task necessitates not just matching the ID label of

Table 2. Comparison with state-of-the-art (SOTA) methods are conducted, considering four different types of Re-ID approaches. The results are obtained using the official released code of these methods on the RAP benchmark.

Method	mAP	rank-1
1.Traditional Re-ID		
EFL [9]	2.01	2.53
LOMO [35]	4.53	7.26
GOG [43]	19.06	31.49
JSTL [58]	14.80	29.94
MSCAN [31]	29.29	48.17
MuDeep [45]	26.26	45.97
HACNN [34]	47.96	70.69
2.Clothing-Change Re-ID		
ReIDCaps [20]	45.40	64.80
CAL [10]	45.50	59.80
3.Re-ID + Attributes		
PAR [36]	38.56	59.76
IDE-ATT-R [32]	38.21	59.25
4.ViT-based architecture		
ViT baseline [5]	45.22	66.18
TransReID baselinee [14]	70.17	82.85
TransReID+CE [14]	70.72	83.21
TransReID+CE+JPM [14]	71.39	83.77
ATPR (Ours)	**74.26**	**86.93**

the image but also aligning with the specified attribute input provided alongside the query image during testing. For a match to be deemed correct, it must correspond to the correct ID and be consistent with the designated attribute label.

4.3. Comparison with SOTA

We first compare our method under the common Re-ID setting that does not specify the attribute of the query image during testing. Tab. 2 shows that in the 'Traditional Re-ID' method section, HACNN achieves high mAP and rank-1 accuracy. In the 'Clothing-Change Re-ID' method section, ReIDCaps and CAL methods score similarly on both metrics, but their performance is slightly worse than HACNN. In the 'Re-ID+Attributes' method, PAR and IDE-ATT-R show high performance even without specific design for intra-class attribute changes, and their mAP and rank-1 metrics exceed some methods in 'Traditional Re-ID' methods. This highlights the role of attribute information in the task. In the 'ViT-based architecture' method section, the TransReID baseline method achieves the highest score on the rank-1 metric, while the TransReID+CE and TransReID+CE+JPM methods also slightly improve the rank-1 metric. Finally, our ATPR achieves the highest mAP and rank-1 metric scores. This indicates that combining attributes, ViT architecture, and Re-ID tasks can achieve better performance on our benchmark.

Tab. 3 presents a comparison of different Re-ID methods under different settings of intra-class attribute changes. Among them, PAR, ReIDCaps, and TransReID are the best-performing Re-ID methods in their respective sections in Tab. 2. In this experiment, we provide the query image along with the desired attributes (the attributes that the query image, with the same ID but different attributes, should possess in the gallery). These attributes are different from the query image's own attributes. Our goal is to retrieve images with the same ID but different attributes. We try assigning attribute changes from 10% to 90% of the queries (s1-s5) and perform retrieval in the gallery. Observing the table, we can note the following points:

- Under all experimental conditions, our ATPR shows higher mAP and rank-1 scores compared to the other three existing methods. This indicates that our method performs well on the AGPR task.
- As the proportion of query images with attribute changes increases, the mAP and rank-1 scores of all methods gradually decrease. Increasing the proportion of query images with attribute changes affects the performance of the AGPR task, which is reasonable.
- Under the same experimental conditions (s1-s5), there are differences in the mAP and rank-1 scores between different methods compared to the results of s0 (non-AGPR task). For example, under s1 (attribute change required), mAP of the PAR method decreases by 4.23% compared to the non-AGPR task (s0), while our method only decreases by 3.93%. This result is based on the fact that our method already has a much higher mAP than PAR under the s0 setting (74.26% vs. 38.56%). Although our method may show slightly higher performance degradation in some conditions, the magnitude of our method's performance decline in s1-s5 (difference between performance decline and non-AGPR performance) is much lower. This highlights the robustness of our method on the AGPR task relative to other methods.

4.4. Ablation Study

The ablation study is provided in Tab. 4:

Impact of different loss combinations: When using only the global loss (L_{gb}) or local loss (L_{loc}), the model's performance is poor. For example, when using 'only L_{glb}', the mAP is 43.78%. This suggests that the global loss ignores important local attribute features. Similarly, using L_{loc} alone disregards the overall correlation among these attributes. Combining both losses can comprehensively leverage their advantages, improving the performance. Adding L_{reg} further constrains the model's training process. For example, with the 'L_{glb}+L_{loc}+L_{reg} (IGAC)' approach (only using IGAC in L_{reg}), the mAP is 48.41%.

Impact of different attribute group combinations: The way attribute groups are combined can influence the model's performance. For instance, merging at-

Table 3. Comparative analyses of methods against varying levels of intra-class attribute (att.) changes. The robustness of each method and our ATPR is evaluated under a spectrum of attribute change scenarios, ranging from no changes (s_0) to 90% alteration (s_1-s_5). The 'no cross att.' setting means we do not specify attributes for retrieval during testing.

Method	s_0:no cross att.		s_1:cross-att.:10%		s_2:cross-att.:30%		s_3:cross-att.:50%		s_4:cross-att.:70%		s_5:cross-att.:90%	
	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1	mAP	rank-1
PAR	38.56	59.76	34.33	55.22	31.35	51.72	27.74	44.35	23.22	37.48	19.57	31.14
$\|s_0\text{-}s_i\|$, i∈[1...5]	0	0	↓4.23	↓4.54	↓7.21	↓8.04	↓10.82	↓15.41	↓15.34	↓22.28	↓18.99	↓28.62
ReIDCaps	45.40	64.80	41.20	60.70	38.60	57.60	30.70	47.50	25.60	36.80	19.20	30.10
$\|s_0\text{-}s_i\|$, i∈[1...5]	0	0	↓4.20	↓4.10	↓6.80	↓7.20	↓14.70	↓17.30	↓19.80	↓28.00	↓26.20	↓34.70
TransReID	71.39	83.77	66.42	77.56	59.35	72.41	52.82	64.28	44.21	55.41	38.91	42.76
$\|s_0\text{-}s_i\|$, i∈[1...5]	0	0	↓4.97	↓6.21	↓12.04	↓11.36	↓18.57	↓19.77	↓27.18	↓28.36	↓32.48	↓41.01
ATPR(Ours)	74.26	86.93	70.33	82.72	66.09	76.41	60.28	70.98	53.24	66.22	49.51	60.33
$\|s_0\text{-}s_i\|$, i∈[1...5]	0	0	↓3.93	↓4.21	↓8.17	↓10.52	↓13.98	↓15.95	↓21.02	↓20.71	↓24.75	↓26.60

Table 4. Ablation Studies

Methods	no cross att.		cross-att.:90%	
	mAP	rank-1	mAP	rank-1
Different Loss Combinations				
only Lglb	70.47	82.82	43.78	53.94
only Lloc	69.98	82.19	42.63	53.76
Lglb+Lloc	71.93	84.02	46.66	57.83
Lglb+Lloc+Lreg (IGAC)	73.17	85.88	48.41	59.10
Lglb+Lloc+Lreg (IGAD)	72.75	85.47	48.67	59.24
Different Attribute Group Combinations				
a^1,(a^2,a^3),(a^4,a^5),a^6	73.26	85.92	48.91	59.26
a^1,a^2,(a^3,a^4,a^5),a^6	72.95	85.71	47.73	58.82
a^1,(a^2,a^3,a^4,a^5),a^6	71.48	84.18	47.65	58.47
Different Patch and Attributes Strategies				
fixed Patch-Att. Assoc.	73.01	85.82	48.46	58.89
w/o. adding AE	68.85	80.49	43.87	55.69
w/o. using [class] in AEG	70.74	83.03	45.22	56.73
Different Inference Strategies				
inference: only f_g	72.92	85.31	48.26	58.83
inference: only $\overline{\sum_{i=1}^{M} f_m^i}$	71.33	84.92	47.95	58.19
ATPR	**74.26**	**86.93**	**49.51**	**60.33**

tribute groups a^2 with a^3, and a^4 with a^5 (i.e., 'a^1,(a^2,a^3),(a^4,a^5),a^6'), results in an mAP of 48.91% and a rank-1 score of 59.26%. Combining a^3, a^4, and a^5 (i.e., 'a^1,a^2,(a^3,a^4,a^5),a^6') yields an mAP of 47.73% and a rank-1 score of 58.82%. This suggests that different attribute group combinations affect the model's understanding of image content. More generalized combinations may lead to a diminished ability to learn specific local attribute information, which in turn can impact the overall performance. For example, amalgamating a^2, a^3, a^4, and a^5 into a single group (i.e., 'a^1,(a^2,a^3,a^4,a^5),a^6') results in lower mAP and rank-1 scores of 47.65% and 58.47%, respectively.

Impact of different patching and attribute strategies: 'fixed Patch-Att. Assoc.': Using fixed patch and attribute association without human keypoint detection prevents the model from dynamically adjusting patch and attribute information associations in different images. This approach achieves an mAP and rank-1 of 48.46% and 58.89%, respectively. 'Without adding AE' 'Without using [class] in AEG': Not using AE or not using [class] after AEG affects the model's performance. For example, 'without adding AE' achieves an mAP and rank-1 of only 43.87% and 55.69%. Not using AE prevents the model from encoding and learning attribute information, impacting its performance. The approach 'without using [class] in AEG' achieves an mAP and rank-1 of 45.22% and 56.73%, respectively, indicating that not using [class] after AEG affects the model's overall understanding of image content.

Different inference strategies: 'inference: only f_g' 'inference: only $\overline{\sum_{i=1}^{M} f_m^i}$': Only using global features or averaging all local information for inference results in poor model performance. For example, the approach 'inference: only f_g' achieves an mAP and rank-1 of only 48.26% and 58.83%, respectively, lacking expression of local information. The approach 'inference: only $\overline{\sum_{i=1}^{M} f_m^i}$' achieves an mAP and rank-1 of only 47.95% and 58.19%, respectively, indicating that using only average attribute features for inference overlooks global pedestrian features.

The last row presents the performance of our AGPR, which achieves the best performance.

4.5. Qualitative Analyses

Fig. 3 demonstrates our ATPR model's capacity for handling attribute-guided pedestrian retrieval. In the first row, without attribute changes, the ATPR model shows proficient baseline matching. The second and third rows present the ATPR's adaptability to attribute changes specified in queries, which is central to the AGPR task. Despite the introduction of new attributes, the ATPR model successfully identifies correct matches, as indicated by the red boxes. This highlights the strength of the local attribute branch and the AEG mechanism in maintaining accurate ID recognition

Figure 3. Rankings with attribute changes. The first row shows initial rankings without specifying desired attributes. Rows two and three show rankings when attributes are modified (in blue) based on witness descriptions. Correct matches are marked with red boxes.

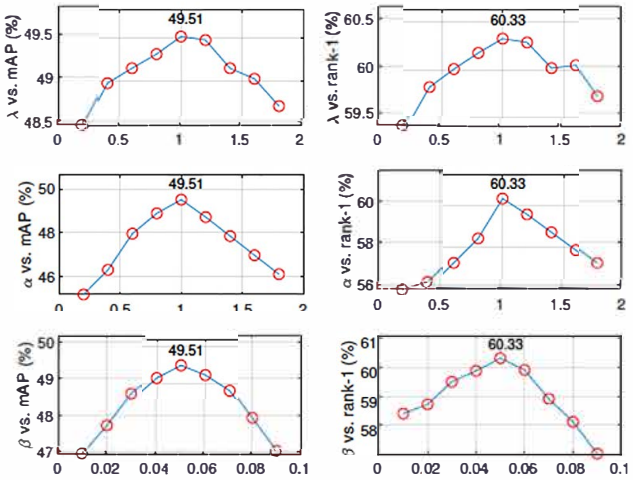

Figure 4. Sensitivity of hyperparameters λ (the first row), α (the second row), and β (the third row.).

amidst attribute variability. Our ATPR's robustness is further underscored by its consistent performance across various attribute alterations, suggesting a well-balanced implementation of IGAC and IGAD within the approach. Overall, this figure substantiates our ATPR's potential in practical AGPR applications, where individuals may alter their appearance yet still need to be reliably identified.

4.6. Hyperparameter Analyses.

The hyperparameter λ in Eq. 6 determines the trade-off between IGAC and IGAD. As observed in Fig. 4, an increase in λ from 0.2 to 1 corresponds with a general increase in both mAP and rank-1 accuracy, suggesting that amplifying the emphasis on IGAC (which highlights inter-group correlations) enhances the performance. At $\lambda = 1$, the model achieves the highest mAP (49.51%) and rank-1 accuracy (60.33%), indicating an optimal balance between intra-group decorrelation and inter-group correlation. A further increase in λ past 1 leads to a decline in performance, implying an overemphasis on inter-group correlations that potentially undermines intra-group decorrelation.

Similarly, α and β modulate the influence of L_{loc} and L_{reg} within the final loss (Eq. 7). As shown in Fig. 4, an upward trend in mAP and rank-1 accuracy is noted as α increases from 0.2 to 1, peaking at α=1 with the highest mAP and rank-1 accuracy, suggesting an ideal balance with the global loss term. Performance drops when α exceeds 1, indicating possible overfitting to local features due to excessive weighting of the local loss component. For β, the mAP and rank-1 accuracy improve consistently as β increases up to 0.05, after which they start to decrease. The peak performance at β=0.05 implies that beyond this point, the regularization may be too stringent, leading to potential model underfitting from excessive attribute correlation/decorrelation.

5. Conclusion

This work presents an Attribute-Guided Transformer-based Pedestrian Retrieval (ATPR) architecture that skillfully merges global ID recognition with local attribute learning. It effectively tackles the challenges posed by the newly introduced Attribute-Guided Pedestrian Retrieval (AGPR) task, specifically person Re-ID with intra-class attribute variability. Our ATPR shows superior performance on the proposed RAP benchmark. However, potential limitations might emerge in scenarios featuring unseen attributes, or significant occlusions, all of which merit further investigation. Future research directions may include extending ATPR to a broader range of datasets and real-world conditions, as well as incorporating additional modalities, such as biometric or behavioral data, to enhance retrieval precision.

6. Acknowledge

This work was jointly supported National Science and Technology Major Project (2022ZD0117901), National Natural Science Foundation of China (62306311, 62373355, 62236010, and 62201061), Fellowship of China Postdoctoral Science Foundation (2022T150698), International Postdoctoral Exchange Fellowship Program of China (YJ20210324), and Special Research Assistant Program of Chinese Academy of Sciences (E2S9180301).

References

[1] Surbhi Aggarwal, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Text-based person search via attribute-aided matching. In *WACV*, pages 2617–2625, 2020. 2, 3

[2] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, pages 7708–7717, 2018. 2

[3] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. *arXiv preprint arXiv:2308.10045*, 2023. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4, 6

[6] Deng-Ping Fan, Ziling Huang, Peng Zheng, Hong Liu, Xuebin Qin, and Luc Van Gool. Facial-sketch synthesis: a new challenge. *Machine Intelligence Research*, 19(4):257–287, 2022. 2

[7] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Fangzhou Lin, Xing Sun, and Xiang Bai. Conditional feature learning based transformer for text-based person search. *IEEE TIP*, 31:6097–6108, 2022. 2, 3

[8] Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *CVPR*, pages 14105–14115, 2022. 2

[9] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 6

[10] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, pages 1060–1069, 2022. 2, 3, 6

[11] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *NeurIPS*, 2018. 2

[12] Devansh Gupta, Aditya Saini, Sarthak Bhagat, Shagun Uppal, Rishi Raj Jain, Drishti Bhasin, Ponnurangam Kumaraguru, and Rajiv Ratn Shah. A suspect identification framework using contrastive relevance feedback. In *WACV*, pages 4361–4369, 2023. 2, 3

[13] Ke Han, Yan Huang, Shaogang Gong, Liang Wang, and Tieniu Tan. 3d shape temporal aggregation for video-based clothing-change person re-identification. In *ACCV*, pages 2371–2387, 2022. 3

[14] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022, 2021. 3, 4, 5, 6

[15] Weizhen He, Shixiang Tang, Yiheng Deng, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Retrieve anyone: A general-purpose person re-identification task with instructions. *arXiv preprint arXiv:2306.07520*, 2023. 3, 5

[16] Han Huang, Yan Huang, and Liang Wang. Vi-diff: Unpaired visible-infrared translation diffusion model for single modality labeled visible-infrared person re-identification. *arXiv preprint arXiv:2310.04122*, 2023. 1

[17] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE TIP*, 28(3):1391–1403, 2018. 1

[18] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2019. 2

[19] Yan Huang, Qiang Wu, JingSong Xu, and Yi Zhong. Sbsgan: Suppression of inter-domain background shift for person re-identification. In *ICCV*, pages 9527–9536, 2019. 1

[20] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE TCSVT*, pages 3459–3471, 2019. 6

[21] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, pages 11895–11904, 2021. 3

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

[23] Yan Huang, Zhang Zhang, Qiang Wu, Yi Zhong, and Liang Wang. Enhancing person re-identification performance through in vivo learning. *IEEE TIP*, 2023. 1

[24] Yan Huang, Qiang Wu, Zhang Zhang, Caifeng Shan, Yi Zhong, and Liang Wang. Meta clothing status calibration for long-term person re-identification. *IEEE TIP*, 2024. 2

[25] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *CVPR*, pages 1629–1636, 2014. 5

[26] Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. Masked vision-language transformer in fashion. *Machine Intelligence Research*, 20(3):421–434, 2023. 2

[27] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797, 2023. 3

[28] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, pages 14278–14287, 2022. 2, 3

[29] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *ICCV*, pages 297–304, 2013. 2

[30] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980, 2012. 2

[31] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, pages 384–393, 2017. 6

[32] Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. *IEEE TIP*, 28(4):1575–1590, 2018. 1, 2, 5, 6

[33] Hongchao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Attribute and state guided structural embedding network for vehicle re-identification. *IEEE TIP*, pages 5949–5962, 2022. 3

[34] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, pages 2285–2294, 2018. 6

[35] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 6

[36] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *PR*, 95:151–161, 2019. 2, 3, 6

[37] Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, and Xiaoming Liu. Learning clothing and pose invariant 3d shape representation for long-term person re-identification. In *ICCV*, pages 19617–19626, 2023. 2, 3

[38] Yunfan Liu, Qi Li, Qiyao Deng, Zhenan Sun, and Ming-Hsuan Yang. Gan-based facial attribute manipulation. *IEEE TPAMI*, 2023. 2, 3

[39] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *ICCV*, pages 2125–2134, 2021. 2

[40] Andong Lu, Zhang Zhang, Yan Huang, Yifan Zhang, Chenglong Li, Jin Tang, and Liang Wang. Illumination distillation framework for nighttime person re-identification and a new benchmark. *IEEE TMM*, 2023. 1

[41] Haoyu Lu, Yuqi Huo, Mingyu Ding, Nanyi Fei, and Zhiwu Lu. Cross-modal contrastive learning for generalizable and efficient image-text retrieval. *Machine Intelligence Research*, 20(4):569–582, 2023. 3

[42] Jinghao Luo, Yaohua Liu, Changxin Gao, and Nong Sang. Learning what and where from attributes to improve person re-identification. In *ICIP*, pages 165–169. IEEE, 2019. 2, 3

[43] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 6

[44] Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval. In *CVPRW*, pages 4547–4557, 2022. 2

[45] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. Multi-scale deep learning architectures for person re-identification. In *ICCV*, pages 5399–5408, 2017. 6

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[47] Arne Schumann and Rainer Stiefelhagen. Person re-identification by deep learning attribute-complementary information. In *CVPRW*, pages 20–28, 2017. 3

[48] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, pages 5566–5574, 2022. 2, 3

[49] Patrick Sudowe, Hannah Spitzer, and Bastian Leibe. Person attribute recognition with a jointly-trained holistic cnn model. In *ICCVW*, pages 87–95, 2015. 2

[50] Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. A simple and robust correlation filtering method for text-based person search. In *ECCV*, pages 726–742, 2022. 2, 3

[51] Chufeng Tang, Lu Sheng, Zhaoxiang Zhang, and Xiaolin Hu. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *ICCV*, pages 4997–5006, 2019. 2

[52] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, pages 7134–7143, 2019. 2, 3

[53] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*, pages 6439–6448, 2019. 2

[54] Xiao Wang, Shaofei Zheng, Rui Yang, Aihua Zheng, Zhe Chen, Jin Tang, and Bin Luo. Pedestrian attribute recognition: A survey. *PR*, 2022. 2

[55] Junyi Wu, Yan Huang, Qiang Wu, Zhipeng Gao, Jianqiang Zhao, and Liqin Huang. Dual-stream guided-learning via a priori optimization for person re-identification. *Transactions on Multimedia Computing, Communications, and Applications*, 17(4):1–22, 2022. 1

[56] Junyi Wu, Yan Huang, Min Gao, Zhipeng Gao, Jianqiang Zhao, Huiji Zhang, and Anguo Zhang. A two-stream hybrid convolution-transformer network architecture for clothing-change person re-identification. *IEEE TMM*, 2023. 3

[57] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. Lapscore: language-guided person search via color reasoning. In *ICCV*, pages 1624–1633, 2021. 3

[58] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016. 6

[59] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 5

[60] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, pages 2029–2046, 2019. 2

[61] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, pages 1472–1481, 2023. 2, 3

[62] Yue Yao, Liang Zheng, Xiaodong Yang, Milind Naphade, and Tom Gedeon. Simulating content consistent vehicle datasets with attribute descent. In *ECCV*, pages 775–791. Springer, 2020. 3

[63] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE TPAMI*, pages 2872–2893, 2021. 1

[64] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, pages 799–807, 2016. 2

[65] Alireza Zaeemzadeh, Shabnam Ghadar, Baldo Faieta, Zhe Lin, Nazanin Rahnavard, Mubarak Shah, and Ratheesh Kalarot. Face image retrieval with attribute manipulation. In *ICCV*, pages 12116–12125, 2021. 2, 3

[66] Jianfu Zhang, Li Niu, and Liqing Zhang. Person re-identification with reinforced attribute attention selection. *IEEE TIP*, 30:603–616, 2020. 3

[67] Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang, and Xianye Ben. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE TMM*, 23:3562–3576, 2020. 3

[68] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *CVPR*, pages 1520–1528, 2017. 2

[69] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 1

[70] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[71] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 1

[72] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, pages 13001–13008, 2020. 5

[73] Chaoran Zhuge, Yujie Peng, Yadong Li, Jiangbo Ai, and Junru Chen. Attribute-guided feature extraction and augmentation robust learning for vehicle re-identification. In *CVPRW*, pages 618–619, 2020. 3

[74] Xiao-Long Zou, Tie-Jun Huang, and Si Wu. Towards a new paradigm for brain-inspired computer vision. *Machine Intelligence Research*, 19(5):412–424, 2022. 2