RA-GAR: A Richly Annotated Benchmark for Gait Attribute Recognition

Chenye Wang[1], **Saihui Hou**[1,2*], **Aoqi Li**[1], **Qingyuan Cai**[1], **Yongzhen Huang**[1,2*]

[1]School of Artificial Intelligence, Beijing Normal University
[2]WATRIX.AI
chenye.wang@mail.bnu.edu.cn, housaihui@bnu.edu.cn, {rookie, caiqingyuan}@mail.bnu.edu.cn,
huangyongzhen@bnu.edu.cn

Abstract

Gait attracts growing interest from researchers due to its advantages as a non-invasive and non-cooperative biometric feature. Current gait-based attribute recognition methods primarily focus on estimating attributes such as gender, age, and emotions. However, there is insufficient attention to diverse gait attributes in various covariate scenarios. In this paper, we design and collect a **R**ichly **A**nnotated benchmark for 15 gait attributes, named **RA-GAR**, comprising data from 533 individuals with over 120,000 sequences. To our knowledge, RA-GAR represents the largest and most diverse benchmark of gait attributes currently available. Furthermore, to fully leverage the semantic information and enhance attribute-specific local perception, we propose a two-stage **CLIP**-based method for **G**ait **A**ttribute **R**ecognition, named **CLIP-GAR**. Experiments on the RA-GAR and MA-Gait datasets demonstrate the effectiveness of CLIP-GAR, showing significant improvements in mean accuracy and F1 score.

Datasets — https://github.com/BNU-IVC/RA-GAR

Introduction

Gait is one of the most promising behavioral biometrics because it can be captured from a distance without requiring any cooperation from the subject (Bouchrika et al. 2011). Current gait attribute recognition methods primarily focus on gender and age attributes, facilitating various applications in surveillance videos (*e.g.*, finding lost children/elderly) (Zhang, Wang, and Li 2022). Additionally, some studies utilize gait to analyze individual emotions, such as happiness, anger, *etc.*, which is considered an important and challenging problem in computer vision and psychology (Hu et al. 2017; Liu et al. 2024).

However, gait as a behavioral biometric encompasses richer attribute information, including toe-in, toe-out, arm swing, *etc.*. Exploring attribute recognition complements the gait community and deepens the understanding of gait patterns. First, the task of attribute recognition contributes to the development of more intelligent gait recognition systems. In the process of tracking criminals, attributes can serve as clues to help security departments quickly track suspects. Second, gait attributes have the potential to enhance the interpretability of gait recognition systems. Mainstream deep learning-based gait recognition methods have proven effective in capturing complex spatio-temporal features, thereby enhancing recognition accuracy (Lin, Zhang, and Yu 2021; Fan et al. 2023a). Nevertheless, the outputs from these black-box models often suffer from a lack of transparency (Chen et al. 2021). Incorporating attribute recognition holds promise for understanding the differences between the probe and gallery.

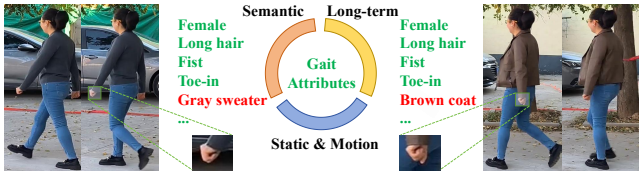


Figure 1: The overview of gait attributes. The attributes proposed in green are semantic attributes with long-term identity invariance and cover both motion (such as toe-in) and static (such as fist) attributes. However, the attributes in red do not meet the long-term identity invariance standards, so they are not within the scope of our consideration.

Most existing open-source gait datasets are primarily designed for identity recognition tasks and only include basic attribute annotations. For instance, CASIA-B (Yu, Tan, and Tan 2006) provides annotations for gender and height attributes. There are a limited number of datasets specifically created for the gait attribute recognition task. Song *et al.* (Song et al. 2023) propose a dataset with multi-attribute annotations tailored for the gait attribute recognition task, named MA-Gait. This dataset comprises recordings of 95 subjects, resulting in over 13,000 sequences. However, MA-Gait does not consider attribute recognition in various covariate scenarios, particularly when subjects are carrying backpacks or changed their clothes, limiting the potential applications of attribute recognition in real-world scenarios. Additionally, the attributes proposed in MA-Gait are relatively limited, with emotional attributes being insufficient for long-term identity recognition.

Therefore, we design and collect a richly annotated

*Corresponding author

benchmark for gait attributes, named RA-GAR, that encompasses view changes, carrying backpacks, dressing variations, and illumination changes. We meticulously craft gait attributes, which are high-level features with *explicit semantic information* and *long-term identity invariance*, encompassing both *static and motion* attributes, as Figure 1 shows. Based on these three criteria, we define 15 gait attributes, making RA-GAR the richest dataset with attributes annotated. Moreover, RA-GAR includes more than 120,000 video sequences involving 533 subjects, establishing it as the largest dataset for gait attribute recognition to our knowledge.

Gait attribute recognition is a classic multi-task, multi-class task, in which an individual possesses numerous attributes simultaneously, with each attribute comprising multiple classes. Existing approaches typically employ a *multi-task classification paradigm*, where a visual backbone is used to extract attribute-related features, and a set of classification heads predict the attribute probabilities (Song et al. 2023). However, these approaches have several limitations: 1) They treats each attribute equally or sequentially according to a predefined order, **ignoring the inherent semantic correlations between attributes**, which have been shown to be effective (Huang et al. 2024) (*e.g.*, a person with humpback often tends to have their head tilted forward.). 2) The model **lacks perception of attribute-specific local regions**. For instance, for head-related attributes, the model should focus on the head region, and for leg-related attributes, it should focus on the lower body. However, in traditional methods, it is challenging to locate these regions, especially when the input is a gait silhouette.

With the rise of multimodal models (*e.g.*, CLIP (Radford et al. 2021)), language models are believed to provide semantic correlation guidance for recognition tasks (Zhu et al. 2023a). Specifically, one-hot labels can be transformed into descriptive sentences, allowing the utilization of pretrained model priors. As a result, recognition tasks can be reformulated into *matching paradigms* (Wang, Xing, and Liu 2021). Furthermore, *multimodal fusion paradigms* have proven more effective for attribute recognition tasks. The fusion operation between modalities allows semantic attributes to interact with corresponding visual tokens, enabling the perception of specific local regions (Wang et al. 2024).

However, directly applying CLIP to the gait attribute recognition task is problematic. The pretrained ResNet50 in CLIP initially achieved an F1 score of 33.97% for zero-shot gait attribute recognition on the RA-GAR benchmark. The reasons for this unsatisfactory phenomenon are twofold: 1) The domain gap between RGB images and silhouette sequences (3 channels *vs.* 1 channel, single-image *vs.* sequences). 2) The recognition granularity. Human-centered attribute recognition requires finer-grained perception compared to natural image recognition. Therefore, we propose using a gait-specific visual encoder instead of the pretrained CLIP visual encoder.

Based on the above discussion, we propose a two-stage gait attribute recognition framework named CLIP-GAR. This framework consists of an **Align stage** and a **Fusion stage**. The purpose of the first stage is to align the gait-specific visual backbone with the pretrained CLIP text encoder for the attribute recognition task. Here, we adopt the simple yet effective DeepGaitV2 (Fan et al. 2023a)-like network as the visual backbone, which has shown promising and robust performance in both in-the-wild and in-the-lab datasets. In the Fusion stage, we freeze both the visual and text encoders, using Transformer encoder blocks to fuse the two modalities, with the fused features serving as input to the attribute classifier. The promising results on RA-GAR and MA-Gait (Song et al. 2023) demonstrate the effectiveness of CLIP-GAR.

In summary, our contributions are as follows:

- We propose three criteria for gait attributes and define 15 attributes based on these criteria. Building upon this, we collect a novel dataset named **RA-GAR**. To our knowledge, it currently stands as the largest and richest dataset for gait attribute recognition.
- We propose a two-stage method for gait attribute recognition, named **CLIP-GAR**. In the first stage, gait-specific features are aligned with attribute features. In the second stage, visual-textual fusion features are generated for gait attribute prediction.
- Comprehensive experiments demonstrate the potential of the RA-GAR benchmark and the effectiveness of CLIP-GAR. Compared to the most relevant method (GAR-Net), CLIP-GAR improved the mean accuracy (mA) by 3.39% and the F1 score by 2.85% on the RA-GAR dataset. On the MA-Gait dataset, it achieved a 2.83% increase in mA and a 0.84% improvement in F1 score.

Related Work

Gait Datasets

CASIA-B (Yu, Tan, and Tan 2006) is one of the most widely used cross-view gait recognition datasets, comprising 124 subjects and three different walking conditions. An increasing number of in-the-wild datasets have been proposed to simulate real-world scenarios. The GREW (Zhu et al. 2021) and Gait3D (Zheng et al. 2022) datasets are constructed from surveillance videos collected in unconstrained environments. In addition to datasets used for identity recognition, there are also open-source datasets focused on gait attributes. OU-ISIR LP-Age (Xu et al. 2017) is introduced to address gait-based age estimation. The INIT Gait Database (Ortells et al. 2018) presents eight walking styles that mimic both typical and pathological gait patterns, facilitating the study of abnormal gait. MA-Gait (Song et al. 2023) focuses on several gait attributes such as toe in, toe out, limping, *etc.*. However, it does not provide gait attributes under different walking conditions, especially in clothes-changing scenarios, limiting the potential applications of attribute recognition in diverse and complex environments. We provide detailed statistics of these datasets in Table 2. [1]

[1]Related work on gait recognition are also provided in the Supplementary materials.

Gait Attribute Recognition

Most existing works on gait attribute recognition can be grouped into three main categories based on their objectives. The first category involves gender and age estimation. Early works on gender and age estimation tasks utilized model-based features, such as leg length and head-to-body ratio (Chuen et al. 2015; Ince et al. 2014). Following the emergence of appearance-based methods, particularly those based on Gait Energy Image (Makihara et al. 2011), an increasing number of researchers adopt machine learning and deep learning approaches to estimating the gender and age of pedestrians (Marín-Jiménez et al. 2017; Xu et al. 2021). Zhang *et al.*. (Zhang, Wang, and Li 2022) introduce a deep ConvNet with multi-task learning, proposing the prediction of a joint distribution instead of two independent distributions. The second category involves emotion recognition algorithms. Li *et al.* (Li et al. 2016) apply the Fourier transform to gait features and utilize statistical methods to identify three different emotional attributes. Bhattacharya *et al.* (Bhattacharya et al. 2020) propose a Spatial Temporal Graph Convolutional Network architecture for classifying human-perceivable emotional attributes. The third category involves recognition algorithms for posture and motion attributes. MA-Gait (Song et al. 2023) is the first to introduce a dataset covering various gait attributes and proposes a Gait Attribute Recognition Network (GAR-Net) to learn static and dynamic information for different attributes.

Datasets

RA-GAR aimed to collect a richly annotated gait dataset with attributes. To comprehensively explore the application potential of gait attributes under various scenarios, we also consider factors such as view angle, clothing, carrying conditions, and illumination changes.

Gait Attribute Definition

Based on the definitions of pedestrian attributes and existing gait attributes (Wang et al. 2022; Song et al. 2023), we summarize and propose the following criteria for gait attributes:

1) Providing clear semantic information. Unlike low-level features derived from statistical measures, the attributes we consider are high-level with explicit semantic meanings. This not only ensures the inherent interpretability of attribute recognition tasks but also establishes the foundation of future downstream tasks based on attribute learning (visual-text alignment (Zhai et al. 2024), attribute editing (He et al. 2019), *etc.*).

2) Possessing long-term identity invariance. We focus on gait attributes that are robust to covariates, excluding emotion-related and subjective attributes (such as beauty, *etc.*) proposed in previous works (Sheng and Li 2021). It is noteworthy that, while numerous pedestrian attributes have been explored in pedestrian re-identification task, most of them focus excessively short-term attributes, such as clothing color and texture (Wang et al. 2022). These attributes lack long-term identity invariance, making them unsuitable for potential applications in clothes-changing condition recognition.

Type		Label	Attribute	#Cls.
Static Attributes	Global	A1	Gender	2
		A2	Age	3
		A3	Height	5
		A4	Body mass	4
	Upper Body	A5	Hair	4
		A6	Forward head position	2
Motion Attributes	Middle Body	A7	Uneven shoulders	3
		A8	Humpback	2
		A9	Hands	4
		A10	Arm swing direction	5
		A11	Arm swing amplitude range	2
		A12	Upper arm movement	2
	Lower Body	A13	Toe	3
		A14	Drag feet	2
		A15	Limping	2

Table 1: The proposed gait attributes. #Cls. represents the number of classes.

3) Both motion and static attributes should be considered. The physical foundation of gait distinctiveness lies in both individual shape differences (*i.e.*, static attributes) and variations in muscle movement patterns during walking (*i.e.*, motion attributes). Previous image-based attributes predominantly emphasized static attributes. However, motion attributes are equally indispensable for gait sequence.

Based on the three criteria, we define 15 gait attributes as outlined in Table 1. Among these, there are 7 binary attributes and 8 multi-class attributes. To our knowledge, RA-GAR provides the richest attribute annotations compared to existing datasets. Further details regarding the attribute definitions can be found in the Supplementary materials.

Data Collection and Processing

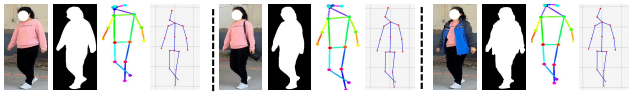
To simulate real-world illumination variations, we set up an outdoor data acquisition platform. Five cameras are fixed at 1.3 meters height to capture from 10 angles (0° to 360° at 36° intervals) with a resolution of 1920×1080 and 30 FPS.

To ensure a rich diversity of attributes and alleviate the issue of class imbalance, each subject is asked to first walk normally along the route and then simulate two combinations of attributes twice following our instructions. Notably, this does not contradict the *long-term identity invariance* characteristic. Essentially, the subjects are simulating pedestrians who exhibit these combinations of gait attributes. Based on the characteristics of the subjects, we customize the attribute combinations to ensure the authenticity of the simulated attributes. For example, we tend to ask elderly subjects to simulate the humpback attribute.

For each attribute combination, subjects are required to walk four times first, then walk twice carrying a bag and finally walk twice after changing clothes. The settings for changing clothes and carrying bags are similar to those in CASIA-B (Yu, Tan, and Tan 2006). In total, each subject is expected to have 240 sequences (10 angles × (1+2) attribute

Dataset	Year	#Id.	#Seq.	#Cam.	Clothes-changing	Gait-oriented attributes	Data types
CAISA-B (Yu, Tan, and Tan 2006)	2006	124	13,640	11	✓	G., H.	Sil., RGB
OU-MVLP (Takemura et al. 2018)	2018	10,307	288,596	14	×	-	Sil.
GREW (Zhu et al. 2021)	2021	26,345	128,671	882	diverse	A., G.	Sil., F., 2/3D pose
Gait3D (Zheng et al. 2022)	2022	4,000	25,309	39	diverse	-	Sil., 2/3D Pose, 3D Mesh
CASIA-E (Song et al. 2022)	2022	1,014	778,752	8	✓	A., G., H., W.	Sil.
OU-ISIR LP-Age (Xu et al. 2017)	2017	63,846	63,846	1	×	A., G.	Sil.
INIT Gait Database (Ortells et al. 2018)	2018	10	160	1	×	8 binary	Sil.
MA-Gait (Song et al. 2023)	2022	95	13,954	6	×	12 binary	Sil., 2D Pose
RA-GAR(ours)	**-**	**533**	**123,067**	**5**	✓	**7 binary, 8 multi-class**	**RGB, Sil., 2/3D Pose**

Table 2: The comparison of RA-GAR with existing gait datasets. #Id. represents the number of identities, #Seq. represents the number of sequences, #Cam. represents the number of cameras, and A., G., H., and W. respectively stand for age, gender, height, and weight. Sil. and F. represent silhouette and optical flow, respectively.



(a) Normal walking (b) Carrying bag (c) Clothing change

Figure 2: Visualization results of RA-GAR. We present the different data modalities of the subjects under three walking conditions, including RGB, silhouette, 2D and 3D pose.

combinations × (4+2+2) walking conditions).

Initially, we manually split the original RGB videos by identity, angles, and various walking conditions. Subsequently, the split clips are processed through a pedestrian tracking network to obtain cropped bounding boxes (Zhang et al. 2022). To fully leverage the application potential of gait attributes across various modalities, PaddleSeg (Liu et al. 2021) is employed to obtain silhouette sequences. For 2D pose estimation, RTMPose (Jiang et al. 2023) and ViTPose (Xu et al. 2022) are utilized to predict 2D keypoints in each frame. Leveraging the output from RTMPose, MotionBert (Zhu et al. 2023b) is used to estimate 3D pose sequences, as illustrated in Figure 2.

Data Statistics and Evaluation Protocols

Ultimately, we collect 533 subjects with 123,067 video sequences after filtering out sequences of low quality. Table 2 summarizes relevant open-source datasets, showcasing our significant advantages in sequence quantity, attribute diversity, covariates factors, and so on. We also present the distribution of sequence length and attributes within RA-GAR in Supplementary materials.

The RA-GAR dataset is divided into two subsets: a training set consisting of 250 randomly selected subjects, and a test set composed of the remaining 288 subjects. The training set comprises 57,155 sequences, while the test set contains 65,912 sequences, both covering the full range of attributes. To assess model performance, we adhere to two well-established evaluation metrics used in gait and pedestrian attribute recognition tasks (Wu et al. 2020; Song et al. 2023; Wang et al. 2022). Firstly, an attribute-based metric known as mean accuracy (mA). Secondly, four instance-based metrics: accuracy (Acc), precision, recall, and the F1 score. The details and formulas of these metrics can be found in the Supplementary materials.

Privacy Statement

We wish to emphasize that all data collection is conducted exclusively for research objectives, adhering strictly to ethical guidelines.All participants involved in data collection signed consent forms and agreed to the release of various data types. We guarantee that the data will be used exclusively for research purposes and that the identity information of the subjects will not be disclosed.

Methods

In this work, we utilize a pretrained textual encoder to model the semantic relationships between gait attributes. We extend attribute annotations into a sentence of attribute description using a unified template. We replace the pretrained visual encoder with a gait-specific encoder, where gait features are horizontally divided into several parts, and each part is pooled into a feature vector. Additionally, we introduce a fusion module that integrates textual features with visual features, facilitating the interaction between attribute features and corresponding part-level visual features.

Assume we have a predefined set of attributes, denoted as $\Pi = \{\{\pi_m^n\}_{n=0}^{N_m}\}_{m=0}^{M}$, where M is the number of attributes, N_m is the number of classes for attribute π_m. We define the total number of attributes class as $\sum_{m=0}^{M}(N_m) = N_{all}$. The training set D_{train} contains I samples $\{(x_i, y_i)\}_{i=0}^{I}$, where $x_i \in \mathbb{R}^{s\times h\times w}$ represents the input gait silhouette, with s as the sequence length, h as the image height, and w as the image width. $y_i \subset \Pi$ denotes the set of human-annotated attribute labels that appear in x_i.

The proposed framework is illustrated in Figure 3. It consists of two training stages: **Align** and **Fusion**. We provide

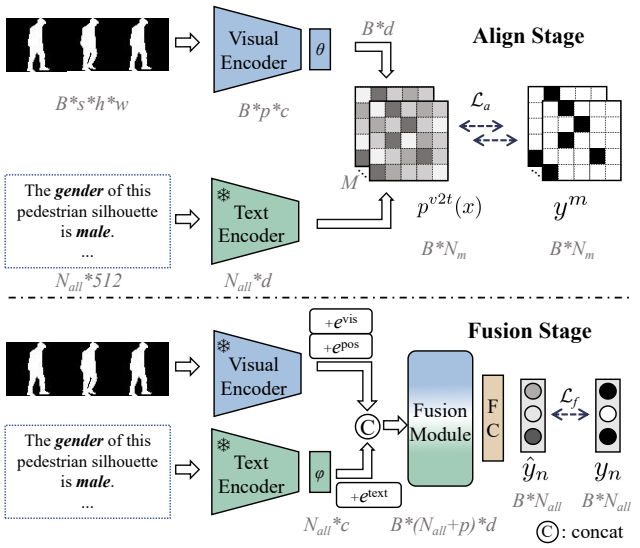


Figure 3: The framework of two-stage CLIP-GAR. In the Align stage, we freeze the parameters of the text encoder. In the fusion stage, the parameters of both the visual and text encoders are kept fixed.

detailed description in the following part.

Align Stage

CLIP provides a powerful pretrained model capable of delivering outstanding performance across various downstream tasks. However, due to the domain gap of silhouettes and RGB images, the potential of CLIP cannot be fully harnessed. Additionally, attribute recognition emphasizes fine-grained differences in human parts and temporal information. To address these challenges, we adopt DeepGaitV2-like network as the gait visual encoder $V(\cdot)$, aligning its features with those of the pretrained textual encoder $T(\cdot)$.

To match the expected textual format of CLIP, we expand the attribute annotation π_m into a sentence of gait attribute description. We design a unified template to expand attribute annotations into sentences: *The* $\{\}$ *of this pedestrian silhouette is/are* $\{\}$. For example, the attribute "male" is expanded to "*The* ***gender*** *of this pedestrian silhouette is* ***male***." The attribute "limping" is expanded to "*The* ***legs*** *of this pedestrian silhouette are* ***limping***." The designed template explicitly reflects the semantic relationships between attributes and locates the attribute-specific regions. The tokenized attribute description $des(\pi_m) \in \mathbb{R}^{N_{all} \times 512}$ is then fed into the frozen pretrained text encoder $T(\cdot)$ to obtain a stable and prior-informed attribute description representation $f_t \in \mathbb{R}^{N_{all} \times d}$, where d is the dimension of the textual representation.

The gait visual encoder $V(\cdot)$ retains the Backbone, Temporal Pooling, and Horizontal Pooling modules from DeepGaitV2, which have been widely validated as effective in gait recognition tasks. The preprocessed silhouette sequence is input into $V(\cdot)$ to extract features $f_v \in \mathbb{R}^{p \times c}$, where p is the number of parts and c is the number of channels.

To match the visual features with the corresponding textual features, a 1x1 convolutional layer θ is applied to reduce the dimension to $\mathbb{R}^d$.

Given a batch of input silhouettes $x \in \mathbb{R}^{B \times s \times h \times w}$, where B is the batch size. For each attributes π_m and corresponding binary ground-truth matrix $y^m \in \mathbb{R}^{B \times N_m}$, to pull the pairwise gait and attribute representations close to each other, we define symmetric similarities between the two modalities with cosine distances:

$$s(x, \pi_m) = \frac{\theta(V(x)) \cdot T(des(\pi_m))^\top}{\|\theta(V(x))\| \|T(des(\pi_m))\|}$$

The softmax-normalized similarity score is:

$$p^{v2t}(x) = \frac{\exp\left(s\left(x, \pi_m^n\right)/\tau\right)}{\sum_{n=1}^{N_m} \exp\left(s\left(x, \pi_m^n\right)/\tau\right)}$$

$$p^{t2v}(\pi_m) = \frac{\exp\left(s\left(\pi_m, x_i\right)/\tau\right)}{\sum_{i=1}^{B} \exp\left(s\left(\pi_m, x_i\right)/\tau\right)}$$

where τ is a learnable temperature parameter. We define the Kullback–Leibler (KL) divergence as the visual-textual contrastive loss to optimize the framework:

$$\mathcal{L}_a = \frac{1}{2}\left\{\mathrm{KL}\left(p^{v2t}(x), y^m\right) + \mathrm{KL}\left(p^{t2v}(\pi_m), (y^m)^\top\right)\right\} \tag{1}$$

Fusion Stage

The purpose of the Fusion stage is to produce visual-textual features from fixed visual features $f_v \in \mathbb{R}^{p \times c}$ and fixed textual features $f_t \in \mathbb{R}^{N_{all} \times d}$ for gait attribute prediction. We use a stack of transformer encoder block $F(\cdot)$ as the fusion module, which achieves cross-modal feature fusion through the self-attention mechanism.

Specifically, we first reduce the dimension of textual features to $\mathbb{R}^{N_{all} \times c}$ via 1x1 convolutional layer ϕ to facilitate concatenation with the visual features. Inspired by (Cheng et al. 2022), we add a learnable model-type embedding e^{text} and e^{vis} to the text and visual features. To encodes spatial parts information into the visual features, we also add a learnable position embedding e^{pos} to the visual features. The concated feature $f_{vt} \in \mathbb{R}^{(N_{all}+p) \times c}$ can be formulated as:

$$f_{vt} = cat\left([\phi(f_t) + e^{\text{text}}, f_v + e^{\text{vis}} + e^{\text{pos}}], dim = 0\right) \tag{2}$$

The concatenated features f_{vt} are subsequently passed through the transformer encoder block $F(\cdot)$, which performs cross-modal feature fusion.

Finally, the predicted attribute logits are generated using a set of attribute classification head $H(\cdot)$ from $F(f_{vt})[: N_{all}, :]$, corresponding to the pretrained textual features f_t.

In the fusion stage, we fixed the parameters of visual encoder $V(\cdot)$ and textual encoder $T(\cdot)$. Binary Cross-Entropy (BCE) loss are used to optimize the fusion module $F(\cdot)$ and classification heads $H(\cdot)$.

$$\mathcal{L}_f = \sum_{n=1}^{N_{all}} \{-[y_n \cdot \log(\hat{y}_n) + (1 - y_n) \cdot \log(1 - \hat{y}_n)]\} \tag{3}$$

where y_n is the ground truth of n-th attributes, $\hat{y}_n = H_n(F(f_{vt})[n, :])$.

Methods		Instance-based metric				Attribute-based metric
		F1	Precision	Recall	ACC	mA
Pose-based	GaitGraph (Teepe et al. 2021)	61.78	77.03	52.34	82.24	56.16
	GaitGraph2 (Teepe et al. 2022)	52.16	58.90	47.77	75.85	50.05
	GaitTR (Zhang et al. 2023)	68.65	73.61	65.01	83.62	62.26
	GPGait (Fu et al. 2023)	69.96	74.97	66.31	84.31	63.93
Silhouette-based	GaitSet (Chao et al. 2019)	70.16	75.72	66.09	84.39	64.33
	GaitPart (Fan et al. 2020)	70.21	74.80	66.81	84.34	63.61
	GaitGL (Lin, Zhang, and Yu 2021)	70.41	**78.04**	64.91	85.03	62.57
	GaitBase (Fan et al. 2023b)	71.07	76.83	66.82	85.06	63.55
	DeepGaitV2-2D (Fan et al. 2023a)	71.22	76.37	67.36	85.00	63.18
	DeepGaitV2-P3D (Fan et al. 2023a)	70.93	76.55	66.71	84.95	61.65
	DeepGaitV2-3D (Fan et al. 2023a)	71.48	76.81	67.44	**85.16**	62.09
	GAR-Net (Song et al. 2023)	73.15	72.65	74.18	84.87	62.18
	CLIP-GAR	**76.00**	77.40	**74.74**	84.31	**65.57**

Table 3: The gait attributes recognition performance on RA-GAR.

Inference

During the inference process, there is no risk of label leakage. All (N_{all}) predefined attribute descriptions are fed into the Text Encoder. These encoded attribute features are then passed into the Fusion Module for prediction. Each attribute has an individual classification head that takes the corresponding feature from the Fusion Module as input. For example, the classification head for "Limping" takes the fused feature derived from the visual feature (from the input silhouette sequence) and the textual feature (from the description, "The legs of this pedestrian silhouette are limping"). It should be emphasized that the text offers specific semantic guidance for the prediction, leading the model to focus on the legs here, regardless of whether the input subject is actually limping.

Experiment

Implementation Details

In our experiments, the input silhouette sequences are standardized to a fixed resolution of 64×44 and a fixed length of 30 frames for fair comparison. We adopt the pretrained CLIP as the textual encoder, with the textual representation dimension d set to 1024. The gait visual encoder follows a DeepGaitV2-3D-like architecture. Following the default configuration, the layers of the backbone are [1,2,2,1], and the channels are [64,128,256,512]. Specifically, the number of bins p in the Horizontal Pooling (HP) is set to 8, with detailed discussions in the ablation study section. The number of transformer encoder blocks is set to 1 in the fusion module.

For both the Align and Fusion stages, we use the Adam optimizer with a learning rate of 10^{-4}, and a weight decay of 2×10^{-5}. The epsilon is set to 10^{-6}. The beta coefficients are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size for both stage is set to 32, with the Align stage training for 120,000 iterations and the Fusion stage training for 20,000 iterations.

Methods	F1	Precision	Recall	ACC	mA
GaitGraph	63.99	72.25	59.63	86.92	67.77
GaitSet	70.81	77.25	67.58	87.62	66.61
GaitPart	69.55	76.65	65.89	87.14	67.87
GaitGL	73.92	79.45	70.87	88.9	67.83
GAR-Net	76.80	77.74	**77.46**	89.34	71.86
CLIP-GAR	**77.64**	**80.92**	75.64	**90.18**	**74.69**

Table 4: The gait attributes recognition performance on MA-Gait.

Gait Attribute Recognition Performance

First, we conduct a comprehensive evaluation of the existing gait-related methods on the RA-GAR benchmark. We compare our approach with four pose-based and six silhouette-based models. We adjust the classification head and loss function of gait recognition methods to accommodate the attribute recognition task.

As claimed in the Datasets section, we adopt label-based and instance-based metrics to evaluate the model performance. Specifically, mA and F1 metrics are more comprehensive for multi-attribute recognition models (Tang et al. 2019; Yang et al. 2021), and CLIP-GAR consistently delivers superior performance. As shown in Table 3, our model achieves state-of-the-art performance across F1, Recall, and mA. We also provide gait attribute recognition performance under different covariates in the Supplementary materials.

CLIP-GAR also generalizes well across different benchmarks. It achieves promising results in MA-Gait, outperforms the second-best model with a margin of mA: 2.83%, F1: 0.84%.

Ablation Studies

Impact of Gait-specific Visual Encoder First, we use the pretrained CLIP visual encoder to extract frame-level gait features, followed by Temporal Pooling to obtain sequence-level features. In the zero-shot setting, CLIP only achieved an F1 score of 33.97%. During the fine-tuning stage, we op-

Visual Encoder		F1	mA
CLIP (ResNet-50)	Zero-shot	33.97	50.08
	Fine-tuned	61.85	50.00
DeepGaitV2-3D	Classification	71.48	62.09
	Stage1 - Align	73.04	62.57
	Stage2 - Fusion	76.00	65.57

Table 5: Performance of different visual encoders in CLIP-GAR.

timized the model using a smaller learning rate (5×10^{-5}). As shown in Table 5, the performance improved noticeably. However, the default input resolution of CLIP is 224×224, which incurs a high computational cost and makes it unfair to compare with mainstream gait methods, which typically use a resolution of 64×44. Therefore, we employed a gait-specific visual encoder and used the Align Stage to obtain pretrained visual and textual features. Compared to directly using the gait-specific visual encoder for classification, the proposed two-stage paradigm more effectively integrates multimodal information, resulting in a significant improvement of 3.48% in mA and 4.52% in F1 score.

Part Numbers	Align Stage		Fusion Stage	
	F1	mA	F1	mA
[16]	72.59	61.66	75.61	65.05
[8]	73.04	62.57	**76.00**	65.57
[4]	**73.67**	**62.67**	75.09	**66.37**
[1]	71.82	61.79	73.25	64.17

Table 6: Influence of different part numbers in CLIP-GAR.

Impact of Different Part Numbers In human-centric recognition tasks, dividing the human body into parts has shown strong discriminative power. Similarly, CLIP-GAR benefits from horizontal feature partitioning. However, the attribute recognition task differs from identity recognition, as the latter focuses on fine-grained distinctions in body parts. We experiment with four different numbers of part divisions, ranging from coarse to fine. As shown in Table 6, both the absence of part divisions and the use of overly fine divisions negatively impact the attribute recognition task. We achieve better results when the number of parts is set to either **8** or **4**.

Discussion and Future Work

Discussion on Identity Recognition

This section discusses the potential impact of gait attributes on identity recognition. First, we assess the performance of identity recognition methods on the RA-GAR dataset. To avoid confusing errors from simulated attributes and to make the dataset more challenging, for each individual, we randomly select seven sequences with real attributes (without distinguishing covariates or viewpoints), designating one as the gallery and using the remaining six as probes. We randomly select 100 subjects for the training set, with the remaining subjects used for testing. We use Rank-1 accuracy and mean Average Precision (mAP) as evaluation metrics. Table 7 presents the recognition performance of five mainstream gait recognition methods on the RA-GAR dataset.

Methods	Without attribute		With attribute	
	Rank-1	mAP	Rank-1	mAP
GaitSet	83.33	88.32	96.11	97.59
GaitPart	76.21	82.96	94.57	96.49
GaitGL	80.56	86.00	83.14	88.10
GaitBase	86.14	90.54	99.54	99.29
DeepGaitV2-P3D	84.99	89.70	96.27	97.65

Table 7: Identity recognition performance of RA-GAR.

During the inference stage, the manually annotated attribute label vector is concatenated along the channel dimension of the gait features. This step can be seen as a way to filter out subjects with inconsistent attributes. Interestingly, all methods exhibit significant performance improvements. Integrating attribute annotations significantly enhances both the performance and interpretability of models. However, accurately predicting gait attributes in real-world applications remains a challenge. RA-GAR enriches the field of gait attribute recognition and provides a valuable data foundation for attribute-assisted identity recognition.

Potential Challenges

One of the challenges is the impact of self-occlusion, which is prevalent in RA-GAR and real-world scenarios. Some fine-grained attributes inevitably suffer from occlusion, presenting a significant challenge in leveraging contextual information for attribute prediction in such obscured scenarios. Additionally, class imbalance is an inherent challenge in attribute recognition tasks. The model needs to be designed to give equal attention to all attributes to mitigate bias. [2]

Conclusion

In this work, we introduce RA-GAR, a dataset of 533 individuals with richly annotated 15 gait attributes, covering view changes, backpacks, clothing, and illumination variations to support gait attribute recognition. To exploit the inherent semantic relationships among attributes and enhance attribute-specific local perception, we propose a two-stage gait attribute recognition method, named CLIP-GAR. We adopt a gait-specific network as the visual encoder, aligning its features with those of the pretrained textual encoder in the first stage. In the second stage, Transform encoders are employed to fuse the cross-modality features. We conduct a comprehensive analysis using gait attribute recognition methods.Comprehensive experiments on RA-GAR and MA-Gait validate the effectiveness of our approach.

[2]Further discussion on gait attributes is provided in the Supplementary Materials.

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China (62206022, 62276025, 62476027) and Beijing Municipal Science & Technology Commission (Z231100007423015).

References

Bhattacharya, U.; Mittal, T.; Chandra, R.; Randhavane, T.; Bera, A.; and Manocha, D. 2020. Step: Spatial temporal graph convolutional networks for emotion perception from gaits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1342–1350.

Bouchrika, I.; Goffredo, M.; Carter, J.; and Nixon, M. 2011. On using gait in forensic biometrics. *Journal of forensic sciences*, 56(4): 882–889.

Chao, H.; He, Y.; Zhang, J.; and Feng, J. 2019. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8126–8133.

Chen, X.; Liu, X.; Liu, W.; Zhang, X.-P.; Zhang, Y.; and Mei, T. 2021. Explainable person re-identification with attribute-guided metric distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11813–11822.

Cheng, X.; Jia, M.; Wang, Q.; and Zhang, J. 2022. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10): 6994–7004.

Chuen, B. K. Y.; Connie, T.; Song, O. T.; and Goh, M. 2015. A preliminary study of gait-based age estimation techniques. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 800–806. IEEE.

Fan, C.; Hou, S.; Huang, Y.; and Yu, S. 2023a. Exploring deep models for practical gait recognition. *arXiv preprint arXiv:2303.03301*.

Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y.; and Yu, S. 2023b. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9707–9716.

Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; and He, Z. 2020. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14225–14233.

Fu, Y.; Meng, S.; Hou, S.; Hu, X.; and Huang, Y. 2023. Gpgait: Generalized pose-based gait recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19595–19604.

He, Z.; Zuo, W.; Kan, M.; Shan, S.; and Chen, X. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11): 5464–5478.

Hu, G.; Hua, Y.; Yuan, Y.; Zhang, Z.; Lu, Z.; Mukherjee, S. S.; Hospedales, T. M.; Robertson, N. M.; and Yang, Y. 2017. Attribute-enhanced face recognition with neural tensor fusion networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 3744–3753.

Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; and Wang, L. 2024. 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*, 17689–17699.

Ince, O. F.; Park, J.; Song, J.; and Yoon, B. 2014. Child and adult classification using ratio of head and body heights in images. *International Journal of Computer and Communication Engineering*, 3(2): 120.

Jiang, T.; Lu, P.; Zhang, L.; Ma, N.; Han, R.; Lyu, C.; Li, Y.; and Chen, K. 2023. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. *arXiv preprint arXiv:2303.07399*.

Li, B.; Zhu, C.; Li, S.; and Zhu, T. 2016. Identifying emotions from non-contact gaits information based on microsoft kinects. *IEEE Transactions on Affective Computing*, 9(4): 585–591.

Lin, B.; Zhang, S.; and Yu, X. 2021. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14648–14656.

Liu, X.; Li, Q.; Hou, S.; Ren, M.; Hu, X.; and Huang, Y. 2024. Depression risk recognition based on gait: A benchmark. *Neurocomputing*, 128045.

Liu, Y.; Chu, L.; Chen, G.; Wu, Z.; Chen, Z.; Lai, B.; and Hao, Y. 2021. Paddleseg: A high-efficient development toolkit for image segmentation. *arXiv preprint arXiv:2101.06175*.

Makihara, Y.; Okumura, M.; Iwama, H.; and Yagi, Y. 2011. Gait-based age estimation using a whole-generation gait database. In *2011 International Joint Conference on Biometrics (IJCB)*, 1–6. IEEE.

Marín-Jiménez, M. J.; Castro, F. M.; Guil, N.; De la Torre, F.; and Medina-Carnicer, R. 2017. Deep multi-task learning for gait-based biometrics. In *2017 IEEE international conference on image processing (ICIP)*, 106–110. IEEE.

Ortells, J.; Herrero-Ezquerro, M. T.; Mollineda, R. A.; and asdfhkl. 2018. Vision-based gait impairment analysis for aided diagnosis. *Medical & biological engineering & computing*, 56: 1553–1564.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sheng, W.; and Li, X. 2021. Multi-task learning for gait-based identity recognition and emotion recognition using attention enhanced temporal graph convolutional network. *Pattern Recognition*, 114: 107868.

Song, C.; Huang, Y.; Wang, W.; and Wang, L. 2022. CASIA-E: a large comprehensive dataset for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 2801–2815.

Song, X.; Hou, S.; Huang, Y.; Cao, C.; Liu, X.; Huang, Y.; and Shan, C. 2023. Gait Attribute Recognition: A New Benchmark for Learning Richer Attributes from Human Gait Patterns. *IEEE Transactions on Information Forensics and Security*.

Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; and Yagi, Y. 2018. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ transactions on Computer Vision and Applications*, 10: 1–14.

Tang, C.; Sheng, L.; Zhang, Z.; and Hu, X. 2019. Improving pedestrian attribute recognition with weakly-supervised multi-scale attribute-specific localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4997–5006.

Teepe, T.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2022. Towards a deeper understanding of skeleton-based gait recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1569–1577.

Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; and Rigoll, G. 2021. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *2021 IEEE international conference on image processing (ICIP)*, 2314–2318. IEEE.

Wang, M.; Xing, J.; and Liu, Y. 2021. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*.

Wang, X.; Zheng, S.; Yang, R.; Zheng, A.; Chen, Z.; Tang, J.; and Luo, B. 2022. Pedestrian attribute recognition: A survey. *Pattern Recognition*, 121: 108220.

Wang, X.; Zhu, Q.; Jin, J.; Zhu, J.; Wang, F.; Jiang, B.; Wang, Y.; and Tian, Y. 2024. Spatio-Temporal Side Tuning Pre-trained Foundation Models for Video-based Pedestrian Attribute Recognition. *arXiv preprint arXiv:2404.17929*.

Wu, J.; Liu, H.; Jiang, J.; Qi, M.; Ren, B.; Li, X.; and Wang, Y. 2020. Person attribute recognition by sequence contextual relation learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10): 3398–3412.

Xu, C.; Makihara, Y.; Liao, R.; Niitsuma, H.; Li, X.; Yagi, Y.; and Lu, J. 2021. Real-time gait-based age estimation and gender classification from a single image. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 3460–3470.

Xu, C.; Makihara, Y.; Ogi, G.; Li, X.; Yagi, Y.; and Lu, J. 2017. The OU-ISIR gait database comprising the large population dataset with age and performance evaluation of age estimation. *IPSJ Transactions on Computer Vision and Applications*, 9(1): 1–14.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35: 38571–38584.

Yang, Y.; Tan, Z.; Tiwari, P.; Pandey, H. M.; Wan, J.; Lei, Z.; Guo, G.; and Li, S. Z. 2021. Cascaded split-and-aggregate learning with feature recombination for pedestrian attribute recognition. *International Journal of Computer Vision*, 129: 2731–2744.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *18th international conference on pattern recognition (ICPR'06)*, volume 4, 441–444. IEEE.

Zhai, Y.; Zeng, Y.; Huang, Z.; Qin, Z.; Jin, X.; and Cao, D. 2024. Multi-Prompts Learning with Cross-Modal Alignment for Attribute-Based Person Re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6979–6987.

Zhang, C.; Chen, X.-P.; Han, G.-Q.; and Liu, X.-J. 2023. Spatial transformer network on skeleton-based gait recognition. *Expert Systems*, 40(6): e13244.

Zhang, S.; Wang, Y.; and Li, A. 2022. Gait Energy Image-Based Human Attribute Recognition using Two-Branch Deep Convolutional Neural Network. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 5(1): 53–63.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, 1–21. Springer.

Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C.; and Mei, T. 2022. Gait recognition in the wild with dense 3d representations and a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20228–20237.

Zhu, J.; Jin, J.; Yang, Z.; Wu, X.; and Wang, X. 2023a. 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*, 2626–2629.

Zhu, W.; Ma, X.; Liu, Z.; Liu, L.; Wu, W.; and Wang, Y. 2023b. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15085–15099.

Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J.; Huang, G.; Du, D.; Lu, J.; and Zhou, J. 2021. Gait recognition in the wild: A benchmark. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14789–14799.