Advances in Artificial Intelligence and Machine Learning; Research 5 (4) 4819-4836 Received 10-09-2025; Accepted 22-12-2025; Published online 29-12-2025

GaitHead: A Universal Head for Gait Recognition Network

Feng Xing xf@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Fangjun Cui cuifj@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Kangxin Wang wangkx4@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Pengcheng Zheng zpc@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Xiaoyang Li lixy5@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Yinkun Chen
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Mingming Wu
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Lijia Wang wanglj13@hn.csg.cn
Haikou Substation Operation and Maintenance Branch,
Hainan Power Grid Co., Ltd.,
China.

Corresponding Author: Lijia Wang

Copyright © 2025 Feng Xingă, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Feng Xing, et al. GaitHead: A Universal Head for Gait Recognition Network. **Advances in Artificial Intelligence and Machine Learning**. 2025;5(4):266. https://dx.doi.org/10.54364/AAIML.2025.54266

Abstract

Gait pattern is one of the most promising biometrics for practical applications since it can be captured at a long distance without requiring intentional cooperation. It is a feasible solution for human identification in transformer substations of power grid. In the latest literature, the basic components of most appearance-based methods can be straightly concluded into three categories, i) the Backbone to extract gait feature maps, ii) the Neck for spatio-temporal feature aggregation, iii) the Head for metric space mapping. Notably, previous works paid much attention to the Backbone and Neck but ignored the significant Head designs. The majority of works separated holistic representations into multiple partial features, but existing Heads are treating all parts equally. However, we argue that treating all parts equivalent in Head seems counterintuitive since it is natural for humans to distinguish two individuals by considering the most significant patterns rather than paying equal attention to each part. To address these issues, we proposed a universal Head named GaitHead, which can be easily applied in most part-based gait recognition networks to adaptively adjust the cross-samples attentive scores among identifying samples according to the identity, camera viewpoint, and walking condition. Specifically, we make the efforts in two aspects for the GaitHead. On the one hand, the Cross-Part Attention Encoder (CPAE) module integrates global information into the part-level attentive representations to obtain the characteristics of the identity, camera viewpoint, and walking condition for each part. On the other hand, the Part Aware Triplet Loss (PAT Loss) is proposed to supervise the cross-sample attentive scores in a well-designed way. Experiments on two common public databases, CASIA-B and OUMVLP, have fully proved that GaitHead is a plug-and-play Head for most part-based gait recognition networks to improve their performance considerably.

Keywords: Gait recognition; Deep learning; Human identification; Part-based gait representation.

1. INTRODUCTION

Gait recognition aims at utilizing the individual walking pattern from video to identify different persons. Compared with face [1], iris [2] or other biometric modalities, human gait can be easily captured in long-distance conditions without the cooperation of subjects, gait recognition possesses great potential in crime investigation, surveillance systems, and social security, so that it has gained increasing interest in both academics and industries. It can be particularly useful for substation monitoring in power grids, as it assists in identifying individuals and preventing unauthorized personnel from entering restricted areas. Therefore, gait recognition should be considered the most feasible solution for this demand. However, many factors like bag-carrying, coat-wearing, and camera viewpoints in realistic scenes lead to dramatic changes in gait appearance. Therefore, the most significant challenge is to extract invariant and unique representations for recognition.

In the latest literature, many innovative appearance-based methods have been proposed to alleviate the above problems. Generally speaking, most of them can be straightly split into three key components as shown in FIGURE 1(a), i.e., i) the Backbone architecture to extract gait representations, ii) the Neck architecture for spatio-temporal feature aggregation, iii) the Head architecture for specific feature mapping. It is worth noting that previous works paid a lot of attention to the Backbone

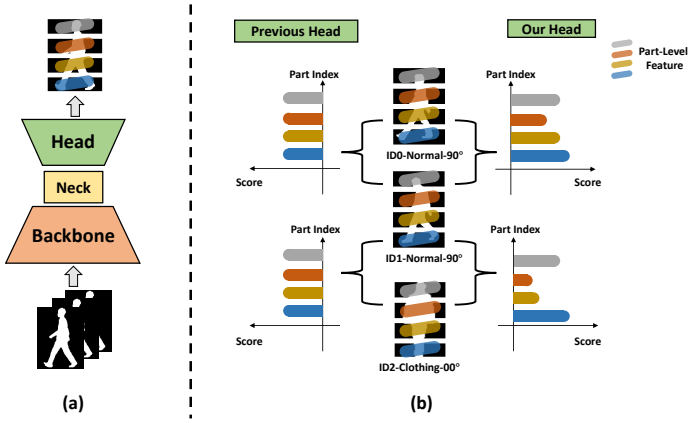

Figure 1: (a): Most appearance-based methods can be straightly split into three key components: i) the Backbone architecture to extract gait representations, ii) the Neck architecture for spatio-temporal feature aggregation, iii) the Head architecture for specific feature mapping. (b): Previous Head VS our Head (GaitHead). The horizontal axis is the value representing the cross-samples attention scores.

and Neck but ignored the significant Head design. From instance, there are various Backbone in recent works, such as the architecture consisted of stacked two dimensional convolutions [3], focal convolutions [4], three dimensional convolutions [5], and 3D local convolutions [6]. The Neck architecture is typically used to capture temporal cues of gait patterns. In other words, it aims to enhance dynamic motion information for gait representations learned from Backbone. The Neck architecture has many variants to extract temporal information such as Set Pooling [3], Micro-motion Capture Module [4], Local Temporal Aggregation [5], and Adaptive Temporal Aggregation [7]. While Horizontal Pooling [3], Salient Spatial Feature Selection [7], Generalized-Mean Pooling [5], and Lateral Pyramid [8], have been exploited for building an spatial representations with effectiveness and discriminativeness. However, to our best knowledge, there are two kinds of heads in our taxonomy, which are the most commonly-used batch normalization neck [5], and compact block [8].

Notably, the majority of works separated holistic representations into multiple partial features, but these existing Heads is treating all parts equally. We argue that average aggregation equally seems counterintuitive since it is natural for humans to distinguish two individuals by considering the most significant patterns rather than paying equal attention to each part. From intuitive human understanding, such adaptive attention on different parts should satisfy three characteristics at least. (1) It should be condition-aware, which means that the model learns to know the clothing condition of candidates for adaptive comparison. (2) It is required to perceive fine-grained differences varying from identities, which refers to identity-aware. (3) The attention needs to become aware of view-invariant representations, which introduces view-aware perception.

In this work, we propose a universal Head (GaitHead) containing three characteristics mentioned above, which can be easily applied in most part-based gait recognition networks to adjust the joint attention between any two samples adaptively. Specifically, GaitHead consists of two novel cooperating modules, the Cross-Part Attention Encoder (CPAE) and the Part Aware Triplet Loss

(PAT Loss). The CPAE module, a special transformer encoder module, firstly takes part-level features as input and then conducts the attentive representation and the distance representation on output. In CPAE, we use the method of self-attention to focus on some crucial parts effectively and combine the part-level attentive representation and global information, including condition, identity, and view information. By this method, the attentive representation from CPAE containing affluent information will make PAT Loss easier to reasonably establish a complete joint probability of any two samples for each part. Finally, the PAT Loss will fine-tune the cross-sample attention score based on join probability according to the appropriate supervision information. Overall, we make the following three main contributions.

- We modified the standard Transformer encoder to create the CPAE, which allows each body part in gait recognition to share global context (like walking conditions and viewpoints) while keeping parts independent.
- We designed the PAT Loss, a simple loss function that uses local and global distance cues to adjust attention scores between samples. It works heuristically to down-weight unreliable parts (e.g., areas blocked by clothing). It can make feature comparisons more adaptive and effective in real-world scenarios.
- We combined the CPAE and PAT Loss into GaitHead, a universal head that can be added to existing gait recognition networks with minimal effort. It consistently boosts performance across different models and conditions, demonstrating practical value as an incremental enhancement.

2. RELATED WORK

In this section, approaches to gait recognition will be generally grouped into two typical categories, e.g., the model-based and appearance-based methods. Besides, the part-based models usually applied in appearance-based methods will be reviewed in short as well.

2.1 Model-Based Methods

This kind of methods [9–11], aims to model the underlying structure of the human body first and then extract the gait motion representations for individual identification. Though robust to some real-world covariation, such as the cross-wearing and bag-carrying conditions, it is still difficult for these methods to precisely model the structure of the human body taken from low-resolution videos.

Some recent works combine silhouettes with poses and can take advantage of different modalities [12]. Optical flow can also be employed to model human body motion in detail [13]. The method SkeletonGait (2024) in [14], shows that the skeleton map, a heatmap of joint coordinates, can achieve state-of-the-art performance.

2.2 Appearance-Based Methods

The appearance-based methods aim at extracting the gait features from silhouettes directly and thus can work in low-resolution conditions [15–17]. According to the input class, these methods can be roughly divided into two categories: template-based and sequence-based. Practically, the input materials of these two kinds of methods are aligned silhouettes sequence. The first one compresses them into one template along the temporal dimension, and the latter regards them as an ordered sequence or an unordered set. Above all, because of being relatively efficient and effective, the development of appearance-based methods presents a more prospering vision in the latest literature.

Beyond, employing partial features for gait description has been more and more prevailing in appearance-based methods. Horizontally splitting the extracted feature map into several pre-defined strips first, and then pooling, embedding them into the metric space, the part-based schemes have proven beneficial for human-centered identification tasks [18–21]. By omitting the spatial alignment, they assume each part-level expression can represent a specific corresponding part of the human body [4, 18, 21]. Lately, Fan et al. (2020) [4], proposed that part-based schemes for gait recognition should be designed in a part-independent manner, owing to the dramatic difference of appearance shape and moving patterns over the human body. Moreover, Qin et al. (2021) [21], considered that the dependency information of each part should be critical, so the feature map cannot be cut with a single fixed scale.

At the same time, we find that the study of gait recognition network Head attracts relatively more minor attention. Specifically, in GaitSet (2019) [3], and GaitPart (2020) [4], their Head is a simple separate fully connected layer (SFC layers). Furthermore, Hou et al. (2020) [8], propose a Compact Block in the Head of GLN [8], to reduce the dimension of the representations. Meanwhile, batch normalization layer and cross-entropy loss are introduced to speed up the convergence and improve the generalization ability of the model in the Head of GaitGL [5]. However, we find these heads treat each part of the part-informed feature equally, vulnerable to attack, especially in cross-clothing conditions. Overall, we design a universal and innovative Head for most part-based gait recognition networks to give them the ability to identify the adaptive joint attention between identifying samples.

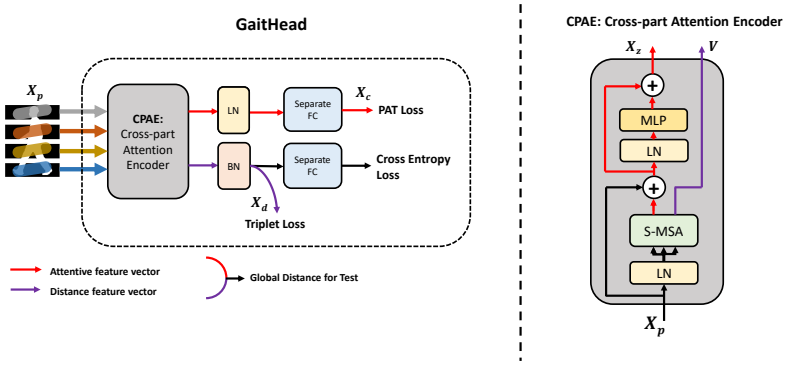

Figure 2: GaitHead Overview. The GaitHead is mainly composed of three parts: i) the Cross-Part Attention Encoder. ii) the PAT Loss to train the attentive feature. iii) the Triplet Loss and the Cross Entropy Loss to train the distance feature like GaitGL [5]. The illustration of the Cross-Part Attention Encoder was inspired by Dosovitskiy et al. (2020) [22].

Similar to our motivation, Wu et al. (2021) [23], proposed a condition-aware comparison scheme for gait recognition to adjust the part weight for each part on different pair conditions. Different from that work, GaitHead can achieve the goal by only modeling Head so that it can be easily transferred to other state-of-the-art methods and demonstrates better robustness.

3. OUR APPROACH

This section proposes a universal Head named GaitHead for part-based gait recognition networks. The pipeline and main components of the GaitHead will be introduced, including the Cross-Part Attention Encoder (**CPAE**) and the Part Aware Triplet Loss (**PAT Loss**). Eventually, the training and testing strategy is presented.

3.1 Pipeline

The overall pipeline of GaitHead is illustrated in FIGURE 2, and it can be easily inserted to these mainstream methods, i.e., GaitSet [3], GaitPart [4], and GaitGL [5], in following four steps.

First, assume $X \in \mathcal{R}^{C\times T\times H\times W}$ is the input of the network representing the sequence of gait silhouettes, where C and T respectively denotes the number of channels and the length of the gait sequence while (H, W) denotes the image resolution. The X can be fed into any desired Backbone, e.g., that of GaitSet [3], GaitPart [4], GaitGL [5], and so on, to extract the frame-level spatial expressions. We formulate this process as $X_b = \text{BackBone}(X)$, where the BackBone denotes the Backbone you want, and $X_b \in \mathcal{R}^{C\times T\times H\times W}$ denotes the sequence of frame-level feature maps.

Next, to gather the part-level understanding of human gait X_p, the sequence X_b will be compressed by the Temporal Pooling (TP) and Horizontal Pooling (HP) module in the Neck, where the order of this two modules is interchangeable and according to the used Backbone. For example, TP is in front of HP in [3, 5], while TP is behind HP in [4]. It is worth noting that we recommend using the Generalized-Mean pooling layer (GeMHPP) block proposed by GaitGL [5], in the HP module. The specific formula is as follows:

$$X_p = \text{GeMHPP}(\text{TP}(X_b)) \qquad X_p \in \mathcal{R}^{N_p\times C} \tag{1}$$

where N_p is the number of parts and X_p denotes the part-level feature representing the corresponding human body part with omitting the spatial alignment. Note that both the partition and pooling strategy here are commonly implemented in the part-based methods [3–5, 19].

After that, we put the part-level feature X_p into GaitHead to extract the distance feature X_d and attentive feature X_c. The first module of GaitHead is the Cross-Part Attention Encoder (CPAE) to transform the part-level feature (X_p) into two features: X_Z and V, which will be mentioned below. Finally, we can get X_c, X_d and the formula is as follows:

$$X_c = \text{LN}(X_Z) \otimes W \quad X_c \in \mathcal{R}^{N_p\times C}, W \in \mathcal{R}^{N_p\times C\times C} \tag{2}$$

$$X_d = \text{BN}(V) \qquad X_d \in \mathcal{R}^{N_p\times C} \tag{3}$$

where LN, BN respectively refers to 1-D layer and batch normalization, and W means a series of separate fully connected layers (SFC layers) while $\otimes$ is a special symbol representing the multiplication operation of this SFC layers.

Ultimately, we will design a reasonable loss function to train these two features: X_c and X_d. We deem highly of the loss designed by GaitGL [5], which is extraordinarily helpful for training the distance feature (X_d). Nevertheless, GaitGL lacks a loss function for training the attentive feature (X_c), so we designed the Part Aware Triplet Loss (PAT Loss). Thus, we select that the Triplet Loss [3–5], and Cross Entropy Loss [5], to train the distance feature (X_d) like GaitGL [5], and the Part Aware Triplet Loss (PAT Loss) designed by us to train the attentive feature (X_c).

3.2 Cross-Part Attention Encoder

3.2.1 Separate multihead self-attention

In CPAE, inspired by Multihead Self-Attention (MSA) [24], we propose a Separate Multihead Self-Attention (S-MSA) block to effectively pay more attention to some significant parts and ensure each part's independence. Specifically, we replace previous fully connected layers (FC layers) with the separate fully connected layers (SFC layers). The formula is as follows and assume that $X \in \mathcal{R}^{N_P \times C}$ is the input of S-MSA:

$$[Q, K, V] = X \otimes W_{qkv} \qquad W_{qkv} \in \mathcal{R}^{N_P \times C \times 3C} \tag{4}$$

$$\mathrm{SA}(X) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_K}})V \tag{5}$$

$$\text{S-MSA}(X) = [\mathrm{SA}_1(X); \mathrm{SA}_2(X); \ldots; \mathrm{SA}_k(X);] \otimes W_s \qquad W_s \in \mathcal{R}^{N_P \times C \times C} \tag{6}$$

where d_K is the dimensions of K. It is worth noting that we will derive V here as the distance feature.

3.2.2 Cross-part attention encoder

The Cross-Part Attention Encoder (CPAE) aims to combine the global spatial feature and the part-level attentive feature so that the identity, camera viewpoint, and condition information of the character can be obtained by each part. Its specific components are shown in FIGURE 2.

The inspiration for creating the CPAE module comes from the Transformer Encoder module in VIT [22], and its principal differences are as follows. First, the previous input is multiple embedded patches, but ours is multiple part-level features X_p without position embedding. Later, the Separate Multihead Self-Attention (**S-MSA**) proposed by us is also different from the previous multi-self attention block (**MSA**). The final difference is that we multiplex the V in S-MSA for calculating the distance feature, and two separate fully connected layers (SFC layers) and the ReLU activations are applied to the MLP module. Therefore, X_Z for indicating attentive feature and V for expressing distance feature are obtained, and the formula is as follows:

$$X_{Z'} = \text{S-MSA}(\mathrm{LN}(X_p)) + X_p \tag{7}$$

$$X_Z = \mathrm{MLP}(\mathrm{LN}(X_{Z'}) + X_{Z'} \tag{8}$$

Overall, CPAE replaces all the Fully Connected layers (FC layers) in Transformer Encoder [22], with the separate fully connected layers (SFC layers) to ensure the independence of each part and substitute the position embedding. The S-MSA module in CPAE effectively pays more attention to some significant parts and reuses the V as the distance feature for measurement. The reusing can reduce the computational cost. Eventually, the CPAE successfully realizes the following functions: cross-part information sharing, essential parts positioning, and part independence, and also obtains two output features X_Z for indicating attentive feature and V for expressing distance feature.

3.3 Part Aware Triplet Loss

The Part Aware Triplet Loss (PAT Loss) aims to highlight the cross-sample attention scores of those parts being still relatively reliable under severe attacks, e.g., when someone in the specific pair wears thick clothes to block his trunk, their heads and legs are the reliable part that is easier to distinguish. Practically, we argue that it is the interaction within the identifying pair that jointly results in the mentioned covariant. Hence the distribution of the cross-sample attention scores among human body parts should be designed as the joint probability of the identifying pair $(a, *)$, mentioned below.

An anchor (a), a positive (p), and a negative (n) samples are treated as a triplet. The following formulation will be taken into account, which is made up of a positive pair (a,p) with a negative pair (a,n), and $* \in \{p, n\}$. Moreover, for uniform, the subscripts below represent the index of the part while the superscripts represent the sample or sample pair category.

3.3.1 Local distance function

Let $\Psi_m(\vec{x}, \vec{y}) : \mathbf{R}^C \times \mathbf{R}^C \to \mathbf{R}$ be a metric function measuring the local distance within the identifying pair $(a, *)$. For clarity we use the shortcut notation $d_i^{a,*} = \Psi_m(\vec{d_i}^{\,a}, \vec{d_i}^{\,*})$ to represent i-th part-level local distance within the identifying pair $(a, *)$, where $\vec{d_i}$ comes from $X_d = \left[\vec{d}_1 \ldots \vec{d}_{N_p}\right]^{\mathrm{T}}$. Practically, the distance function Ψ_m is instantiated as the L2 distance, and formulated as $d_i^{a,*} = ||\vec{d_i}^{\,a} - \vec{d_i}^{\,*}||$.

Let $D_{local_i}^{a,p,n} = d_i^{a,p} - d_i^{a,n} + m$ represent local level distance difference between the positive and negative pairs in i-th part, where m denotes the margin and is set to 0.2. By the way, the Triplet Loss function is below:

$$\mathcal{L}_{triplet_i}^{a,p,n} = \lfloor D_{local_i}^{a,p,n} \rfloor_+ \tag{9}$$

,where $\lfloor x \rfloor_+ = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{else} \end{cases}$.

3.3.2 Global distance function

Let $\Psi_c(\vec{x}, \vec{y}) : \mathbf{R}^C \times \mathbf{R}^C \to \mathbf{R}$ be a joint probability function measuring the cross-samples attention scores within the identifying pair. For clarity we use the shortcut notation $\omega_i^{a,*} = \Psi_c(\vec{c_i}^{\,a}, \vec{c_i}^{\,*})$ to represent i-th part-level cross-samples attention score within the identifying $pair(a, *)$, where $\vec{c_i}$

comes from $X_c = \left[\vec{c}_1 \ldots \vec{c}_{N_p}\right]^{\mathrm{T}}$. Practically, the function Ψ_c is instantiated as the cosine function plus 1 and then divide 2, and formulated as $\omega_i^{a,*} = \frac{Cos(\vec{c}_i^{\,a}, \vec{c}_i^{\,*})+1}{2}$, where Cos denotes the cosine distance.

The global-understanding distance within the identifying pair is defined as the weighted-mean of the local part-level distances, and formulated as $G^{a,*} = \frac{\sum_i^{N_p} d_i^{a,*} \omega_i^{a,*}}{\sum_i^{N_p} \omega_i^{a,*}}$, where $d_i^{a,*}$ and $\omega_i^{a,*}$ respectively represents i-th part-level local distance and cross-samples attention score within the identifying $pair(a, *)$.

Let $D_{global}^{a,p,n} = G^{a,p} - G^{a,n} + m$ represent global level distance difference between the positive and negative pairs, where m denotes the global margin and is set to 0.2.

3.3.3 Part aware triplet loss

In order for the Part Aware Triplet Loss (PAT Loss) to effectively train the attentive feature, we intend to use local distance $D_{local_i}^{a,p,n}$, global distance $D_{global}^{a,p,n}$ and $\bar{d}^{a,*}$ as its supervision information, where $\bar{d}^{a,*}$ denotes the mean understanding of distance and has been defined as $\bar{d}^{a,*} = \frac{\sum_{i=1}^{n_p} d_i^{a,*}}{n}$. When the $d_i^{a,p} > \bar{d}^{a,p}$, we consider that the $d_i^{a,p}$ in i-th part may be unreliable. Since pair (a, p) is a positive pair, the distance of i-th part should be small, and it is better not to be greater than the average distance of the whole part. Similarly, when the $d_i^{a,n} < \bar{d}^{a,n}$, we consider that the $d_i^{a,n}$ in i-th part may be unreliable. Of course, one of the preconditions is that when the local triplet loss is generated, $D_{local_i}^{a,p,n} > 0$. Based on the above analysis, we define a signal flag $S_i^{a,*}$, and the formula is as follows:

$$S_i^{a,p} = \begin{cases} 1 & \text{if } D_{local_i}^{a,p,n} > 0 \text{ and } d_i^{a,p} > \bar{d}^{a,p} \\ 0 & \text{else} \end{cases} \tag{10}$$

$$S_i^{a,n} = \begin{cases} 1 & \text{if } D_{local_i}^{a,p,n} > 0 \text{ and } d_i^{a,n} < \bar{d}^{a,n} \\ 0 & \text{else} \end{cases} \tag{11}$$

When $S_i^{a,*} = 1$, we need to reduce the value of the cross-samples attention scores $\omega_i^{a,*}$ in the i-th part under the identifying pair $(a, *)$, because the distance $d_i^{a,*}$ in i-th part is no longer trustworthy. In general, the PAT Loss formula is as follows:

$$\mathcal{L}_{pat_i}^{a,p,n} = \begin{cases} BCE(\omega_i^{a,p}, S_i^{a,p}) + BCE(\omega_i^{a,n}, S_i^{a,n}) & \text{if } D_{global}^{a,p,n} > 0 \text{ and } D_{local_i}^{a,p,n} > 0 \\ 0 & \text{else} \end{cases} \tag{12}$$

where BCE is the Binary Cross Entropy Loss [25], and $BCE(x, y) = -(1 - y)\ln(x) - y\ln(1 - x)$.

On the whole, we reasonably use the relationship between the three critical data ($D_{local_i}^{a,p,n}$, $D_{global}^{a,p,n}$ and $\bar{d}^{a,*}$) as a supervision signal to train the attentive feature, so that the attentive feature can successfully express the cross-sample attention scores based on the joint probability of any two samples.

3.4 Training and Testing Strategy

3.4.1 Training strategy

In the whole GaitHead framework, we have three losses, standard triplet loss $\mathcal{L}_{triplet}^{a,p,n}$ [26], PAT Loss $\mathcal{L}_{pat}^{a,p,n}$ and Cross Entropy Loss $\mathcal{L}_{id}$. In general, the full-term loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{triplet}^{a,p,n} + \lambda \mathcal{L}_{pat}^{a,p,n} + \mathcal{L}_{id} \tag{13}$$

, where λ is set to 0.002. Its value is determined empirically based on validation performance.

Note the supervision information to attentive feature almost entirely from the distance feature, which indicates that the superior distance feature vector drives the attentive feature better.

3.4.2 Testing strategy

The distance between probe and gallery is defined as the global distance $G^{a,*}$, which is the weighted mean of the distance and the cross-sample attention scores between each part of the probe and gallery. Eventually, we use this global distance $G^{a,*}$ to find the nearest neighbor as rank 1 to calculate the accuracy.

4. EXPERIMENTS

4.1 Settings

4.1.1 Database

All the experiments are conducted on two employed public databases, i.e., CASIA-B [27], and OUMVLP [28].

CASIA-B [27], is composed of 124 subjects, and each subject contains 3 walking conditions, i.e., normal cases (NM#1-6), bag-carrying (BG#1-2), and clothes-change (CL#1-2). In addition, each walking condition is taken from 11 views uniformly distributed in $[0°, 180°]$. Hence, there are $11 \times (6 + 2 + 2) = 110$ sequences per subject ideally. In our experiments, the first 74 subjects are grouped into the training set, while the remaining 50 subjects are grouped for evaluation. At test stage, for each subject, the first 4 sequences under the NM condition (NM#1-4) are grouped as the gallery set, and the remaining 6 sequences are divided into 3 probe subsets according to their walking condition, i.e., the NM, BG, and CL subset.

OUMVLP [28], is one of the largest gait datasets in public. This dataset is composed of 10307 subjects, and each subject only contains one walking condition, i.e., normal cases (NM#0-1). Each walking condition is taken form 14 views uniformly distributed between $[0°, 90°]$ and $[180°, 270°]$. Therefore, there are $2 \times 14 = 28$ sequences per subject ideally. Our experiments strictly follow the official protocols and take 5153 subjects as the training set while the rest 5154 subjects as the test

set. At the test stage, the first sequence (NM#0) is regarded as the probe set for each subject, and another sequence (NM#1) is used for the gallery.

4.1.2 Network details

The input/output dimension and the number of parts in GaitHead are as much as possible the same as the origin Head. All LayerNorm layers affine transformation in GaitHead are closed and normalized only to the channel dimension. In the Separate Multihead Self-Attention block (S-MSA), we set the number of headers to 4. Meanwhile, if part-based networks that want to use GaitHead, we will use the GeMHPP block proposed in GaitGL [5], to replace its original HP module. It is worth noting that the X_Z from S-MSA is detached from X_p. Besides, the activation function in MLP is the RELU function.

4.1.3 Implementation details

All models are implemented with PyTorch [29]. All the input silhouettes are aligned by the methods utilized in [30], and resized to the size of 64×44 for CASIA-B and OUMVLP. The GaitHead will not change the optimizer, so the optimizer uses the corresponding optimizer mentioned in the original method, such as Adam [31], for GaitGL and GaitPart and so on. For batch size and learning rate, we will not change it as well and use the batch size mentioned in the paper of the corresponding method. In practice, we first run the original model in the first few iterations and then add the GaitHead to run the remaining iterations. When running the GaitHead, the learning rate will gradually decline. It is worth mentioning that we reproduce the results for GaitSet, GaitPart, and GaitGL in CASIA-B and OUMVLP datasets, and GaitHead also conducted experiments on this basis. Therefore, we put the reproduced results in all the experimental tables below to compare the benefits of GaitHead.

4.2 Performance Comparison

4.2.1 CASIA-B

We add the proposed GaitHead to GaitSet [3], GaitPart [4], GaitGL [5], and MSAFF [32], and compare their performance with the latest gait recognition approaches, including GaitSet [3], GaitPart [4], GLN [8], CSTL [7], GaitGL [5], and MSAFF [32] (TABLE 1). To ensure the comparisons systematical and comprehensive, we take all the cross-view and cross walking conditions cases into account. Overall, the GaitHead mainly improve the GaitSet, GaitPart, GaitGL and MSAFF remarkably. More specifically, GaitHead helped the GaitSet increase by +1.6%, +2.7% and +5.0%, GaitPart increased by +0.5%, +1.3% and −0.1%, GaitGL increased by +1.0%, +1.5% and +3.9%, and MSAFF increased by +0.3%, +0.7% and +1.4% in NM, BG and CL cases respectively. The accuracy of MSAFF plus GaitHead reaches the highest accuracy on various walking conditions, i.e., 99.3%, 97.8%, and 94.7% for NM, BG, and CL, respectively, which outperforms GaitGL itself as well as other methods significantly.

Table 1: Averaged rank-1 accuracies on **CASIA-B**, excluding identical-view cases.

Probe	Method	GaitHead	Probe View											Mean
			0°	16°	32°	54°	72°	90°	108°	126°	154°	162°	180°	
NM	GaitSet 2019	×	92.3	98.3	99.3	97.8	94.4	91.9	95.6	97.5	98.4	98.0	90.1	95.7
		√	94.7	**99.7**	**99.8**	98.6	95.6	95.3	97.2	98.8	99.4	98.8	93.2	97.3 (↑1.6)
	GaitPart 2020	×	93.0	97.9	99.2	98.3	95.0	92.2	96.4	98.3	99.0	97.4	90.7	96.1
		√	94.1	98.6	99.7	98.3	96.0	92.8	96.8	98.4	99.3	97.5	91.6	96.6 (↑0.5)
	GLN 2020	×	93.2	99.3	99.5	98.7	96.1	95.6	97.2	98.1	99.3	98.6	90.1	96.8
	CSTL 2021	×	97.2	99.0	99.2	98.1	96.2	95.5	97.7	98.7	99.2	98.9	96.5	97.8
	TransGait 2023	×	97.3	99.6	99.7	**99.0**	97.1	95.4	97.4	99.1	99.6	98.9	95.8	98.1
	GaitGL 2021	×	96.2	98.5	99.1	98.0	96.6	95.2	97.2	99.0	98.9	98.8	94.0	97.4
		√	96.6	99.0	99.5	98.8	97.9	96.6	98.2	**100.0**	99.6	99.6	96.8	98.4 (↑1.0)
	MSAFF 2024	×	99.1	99.4	99.3	99.1	98.9	98.9	98.9	99.2	**99.7**	99.6	97.8	99.1
		√	**99.2**	99.5	99.5	**99.0**	**99.1**	**99.3**	**99.3**	99.2	99.3	**99.7**	**98.9**	**99.3** (↑0.2)
BG	GaitSet 2019	×	88.0	94.3	94.2	91.9	86.0	82.4	86.1	91.5	95.0	95.6	85.0	90.0
		√	89.2	95.5	96.2	94.2	89.9	85.8	91.6	94.2	97.1	96.6	89.7	92.7 (↑2.7)
	GaitPart 2020	×	88.6	92.5	94.5	93.0	89.0	83.5	87.7	93.0	95.9	93.9	85.5	90.6
		√	90.4	94.9	94.8	93.8	90.3	86.2	89.0	93.7	97.0	95.2	86.6	91.9 (↑1.9)
	GLN 2020	×	91.1	97.6	97.7	95.2	92.5	91.2	92.4	96.0	97.5	94.9	88.1	94.0
	CSTL 2021	×	91.7	96.5	97.0	95.4	90.9	88.0	91.5	95.8	97.0	95.5	90.3	93.6
	TransGait 2023	×	94.0	97.1	96.5	96.0	93.5	91.5	93.6	95.9	97.2	97.1	91.6	94.9
	GaitGL 2021	×	93.3	96.1	97.1	96.3	94.2	88.6	92.7	95.8	98.0	96.4	91.4	94.5
		√	94.8	98.1	96.9	96.4	95.5	93.5	94.9	96.9	98.2	97.9	93.0	96.0 (↑1.5)
	MSAFF 2024	×	97.7	98.5	98.6	98.0	96.9	95.3	96.2	97.6	98.5	97.7	94.1	97.1
		√	**97.9**	**98.8**	**98.9**	**98.5**	**97.5**	**96.9**	**96.3**	**98.2**	**98.6**	**98.1**	**96.1**	**97.8** (↑0.7)
CL	GaitSet 2019	×	68.9	82.7	85.4	80.4	73.1	69.5	73.3	76.7	77.9	79.1	62.7	75.4
		√	74.8	87.8	88.2	84.1	78.9	75.9	80.0	82.2	83.4	81.1	68.3	80.4 (↑ 5.0)
	GaitPart 2020	×	72.2	84.3	86.5	82.3	77.3	74.2	77.9	80.0	84.0	80.3	66.5	78.6
		√	70.6	84.5	87.8	82.4	77.9	73.4	77.2	79.5	82.8	81.4	66.8	78.5 (↓ 0.1)
	GLN 2020	×	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	CSTL 2021	×	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	75.3	84.2
	TransGait 2023	×	80.1	89.3	91.0	89.1	84.7	83.3	85.6	87.5	88.2	88.8	76.6	85.8
	GaitGL 2021	×	77.1	90.4	92.1	90.1	82.7	77.7	81.8	86.2	88.8	84.6	70.6	83.8
		√	79.6	93.6	93.9	90.9	88.6	84.0	87.9	90.4	92.3	88.9	75.1	87.7 (↑ 3.9)
	MSAFF 2024	×	92.1	94.6	95.6	93.8	91.0	90.6	92.5	94.0	95.3	94.8	91.7	93.3
		√	**93.8**	**95.4**	**96.1**	**95.2**	**93.2**	**93.7**	**93.1**	**94.8**	**95.9**	**96.0**	**94.3**	**94.7** (↑ 1.4)

Regarding the performance of GaitPart integrated with GaitHead, we observe that the improvement under cross-clothing (CL) conditions is not as obvious as with other models, such as GaitSet and GaitGL. This limited enhancement may stem from a suboptimal interaction between GaitHead's hyper-parameters and GaitPart's architectural design, as GaitPart is relatively sensitive to experimental configurations, with similar instances noted in prior studies [33, 34]. Despite this specific case, the collective results demonstrate that GaitHead consistently enables models like GaitSet, GaitPart, GaitGL, and the recently added MSAFF (2024) to effectively aware of the crossing walking conditions and adjust the weights distributed among human body parts.

4.2.2 OUMVLP

As shown in TABLE 2, the performance of MSAFF equipped with GaitHead is also better than MSAFF itself and other methods on the OUMVLP dataset. Specifically, GaitHead boosts the MSAFF by +0.8% on this one of the worldwide largest gait datasets. Since there is no cross walking conditions cases in OUMVLP, we thus claim that the GaitHead provides MSAFF the capacity to

Table 2: Averaged rank-1 accuracies on **OUMVLP**, excluding identical-view cases.

Method	GaitHead	Probe View														Mean
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet 2016	×	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet 2019	×	78.6	87.4	89.8	90.0	87.7	88.4	87.4	81.2	86.2	88.8	89.0	87.1	87.5	86.1	86.8
GaitPart 2020	×	82.1	88.7	90.7	90.8	89.5	89.6	89.1	84.7	87.3	89.8	89.9	88.7	88.8	87.7	88.4
GLN 2020	×	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
CSTL 2021	×	87.1	91.0	**91.5**	**91.8**	90.6	90.8	90.6	89.4	**90.2**	90.5	90.7	89.8	90.0	89.4	90.2
GaitGL 2021	×	84.3	89.8	91.1	91.4	90.8	90.5	90.1	88.2	88.0	90.2	90.3	89.5	89.4	88.6	89.5
	√	**87.4**	**91.2**	**91.5**	91.7	**91.4**	**91.1**	**90.7**	**89.7**	**90.2**	**90.7**	**90.8**	**90.3**	**90.1**	**89.7**	**90.5** (↑ 1.0)

perceive the view changes within the identifying pair and accordingly adjust the weights distributed among human body parts positively. The following visualization experiments will illustrate this conclusion more intuitively.

Table 3: Ablation Study on CASIA-B

Method	GaitHead		Probe			Mean
	CPAE	PAT	NM	BG	CL	
GaitSet 2019	×	×	95.7	90.0	75.4	87.0
	√	×	97.1	92.4	77.9	89.1
	√	√	**97.3**	**92.7**	**80.4**	**90.1**
GaitPart 2020	×	×	96.1	90.6	**78.6**	88.4
	√	×	**96.6**	**91.9**	76.9	88.4
	√	√	**96.6**	**91.9**	78.5	**89.0**
GaitGL 2021	×	×	97.4	94.5	83.8	91.9
	√	×	98.5	95.8	86.5	93.6
	√	√	**98.4**	**96.0**	**87.7**	**94.0**
MSAFF 2024	×	×	99.1	97.1	93.3	96.5
	√	×	99.1	97.5	94.1	96.9
	√	√	**99.3**	**97.8**	**94.7**	**97.3**

4.3 Ablation Study

4.3.1 CPAE and PAT loss

Since PAT loss is not indispensable for GaitHead, we strip it from GaitHead to further analyze the effectiveness of the Cross-Part Attention Encoder (CPAE). Besides, as mentioned above, the GaitHead is a novel head for part-based gait recognition networks. Hence, we equip several popular methods with GaitHead, i.e., GaitSet, GaitPart, and GaitGL, to experimentally show the robustness and superiority of our model. As shown in TABLE 3, the design of the CPAE plus PAT Loss module has achieved the highest average score in these three methods to prove that all components of GaitHead are effective compatible.

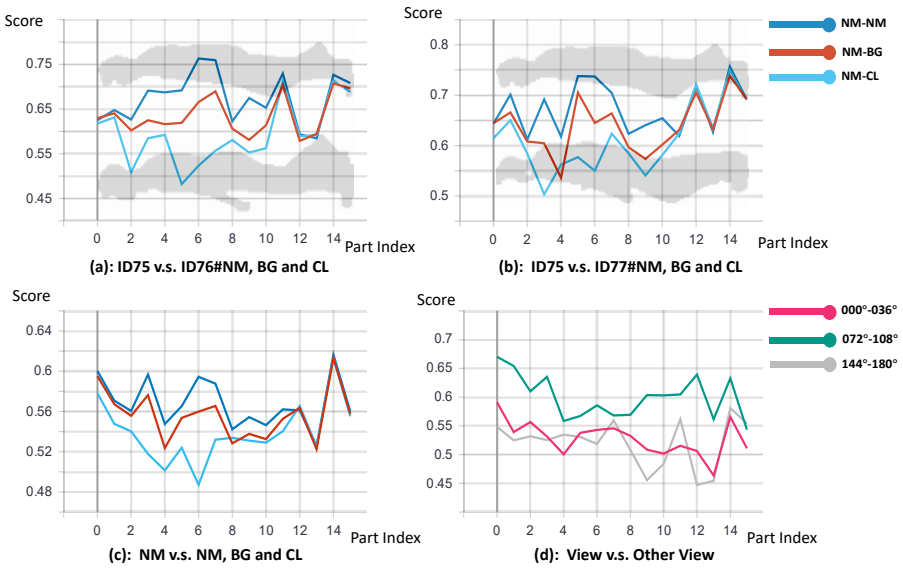

Figure 3: The horizontal axis divides the silhouette into 16 parts from top to bottom. The vertical axis is the value representing the cross-samples attention scores ($\omega_i^{a,*}$), ranging from 0 to 1. Different colored lines represent pairs between probe and gallery under different dress conditions. NM-CL means that probes are in normal (NM) state, while galleries are in clothing (CL) condition. 000°-036° express that probe view is 0° and gallery view is 36°. **The data here comes from GaitSet+GaitHead running on the CASIA-B test set.**

4.3.2 Visualization of the cross-samples attention

In our work, the adaptive joint attention among different samples, i.e., the attention scores based on the joint probability of any two samples, can introduce the condition-aware, identity-aware and view-aware perceptions. To make it intuitive, we conduct the following visualization experiments.

For brevity's sake, here randomly take the GaitSet equipped with GaitHead as an instance, and actually, the experiments based on other methods can reveal similar conclusions as well. As shown in Figure 3, the horizontal represents the index of a human body part (ranged from 0 to 15 v.s. from head to foot, just the first 16 parts are exhibited here for clarity), and the vertical axis represents the cross-samples attentive scores or the average.

The FIGURE 3(a), illustrates that, in a specific identifying pair, i.e., ID75 and ID76, the parts#0-1 above the shoulder and the parts#11-15 below the knee possess relatively stable attentive scores regardless of the walking conditions. But for the rest of parts#3-11, their cross-samples attentive scores consistently decrease with the walking condition varying from the NM into the CL, in other words, with the increasing severity of occlusions over the human torso. Hence, we claim that GaitHead owns the ability to perceive the walking conditions transformation among the identifying samples and is helpful to adjust the cross-samples attentive scores adaptively and positively.

Comparison between FIGURE 3(a) and (b), exhibits that GaitHead perceives fine-grained differences varying from identities as well, owing to the different identifying pairs possessing the various attentive scores distribution over the human body. FIGURE 3(c) and (d), shows the distribution of the cross-samples attentive scores' average over the human body also vary according to the walking condition transformations, i.e., the cross viewpoints and clothing cases.

In summary, we announce that the GaitHead can explore the condition-aware, identity-aware, and view-aware perceptions effectively and correspondingly adjust the distribution of cross-samples attention scores over the human body adaptively and positively.

5. CONCLUSION

In this paper, we put forward a universal Head named GaitHead, which can be easily applied in most part-based gait methods to adaptively adjust the cross-samples attentive scores among identifying samples according to the identity, camera viewpoint, and walking condition. First, to integrate the global information into the part-level feature, we propose the Cross-Part Attention Encoder to ensure part independence and make the attentive representation have a global vision to obtain a person's identity, camera viewpoint, and walking information. Second, to adjust the cross-sample attention scores, we present the Part Aware Triplet Loss to build the joint probability of any two samples. Besides, the experiment conducted on two common databases, i.e., the CASIA-B [27], and OUMVLP [28], adequately show the "plug and play" yet effective GaitHead may help most of the part-based models improve performance. Moreover, further visual experiments ensure our goal is achieved undoubtedly.

6. ACKNOWLEDGMENT

This work was supported in part by *Key Technologies for Unobtrusive Personal Safety Prevention and Control in Substations of Power Grids* (Grant 075000KC23120004).

References

[1] Hofer P, Roland M, Mayrhofer R, Schwarz P. Optimizing Distributed Face Recognition Systems through Efficient Aggregation of Facial Embeddings. Adv Artif Intell Mach Learn. 2023;3:693-711.

[2] Daugman J. Understanding Biometric Entropy and Iris Capacity: Avoiding Identity Collisions on National Scales. Adv Artif Intell Mach Learn. 2024;4:2152-2163.

[3] Chao H, He Y, Zhang J, Feng J. Gaitset: Regarding Gait as a Set for Cross-View Gait Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:8126-8133.

[4] Fan C, Peng Y, Cao C, Liu X, Hou S, et al. Gaitpart: Temporal Part-Based Model for Gait Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition. IEEE/CVF. 2020:14225–14233.

[5] Lin B, Zhang S, Yu X. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF. 2021;14:14628-14636.

[6] Huang Z, Xue D, Shen X, Tian X, Li H, et al. 3D Local Convolutional Neural Networks for Gait Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF. 2021;14:14920-14929.

[7] Huang X, Zhu D, Wang H, Wang X, Yang B, et al. Context-Sensitive Temporal Feature Learning for Gait Recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF. 2021;12:12909-12918.

[8] Hou S, Cao C, Liu X, Huang Y. Gait Lateral Network: Learning Discriminative and Compact Representations for Gait Recognition. In: European Conference on Computer Vision. Springer. 2020:382-398.

[9] Wang Y, Yu S, Wang Y, Tan T. Gait Recognition Based on Fusion of Multi-View Gait Sequences. In: International Conference on Biometrics. Springer. 2006:605-611.

[10] Bodor R, Drenner A, Fehr D, Masoud O, Papanikolopoulos N. View-Independent Human Motion Classification Using Image-Based Reconstruction. Image Vis Comput. 2009;27:1194-1206.

[11] Liao R, Cao C, Garcia EB, Yu S, Huang Y. Pose-Based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations. In: Chinese Conference on Biometric Recognition. Springer. 2017:474-483.

[12] Zhao Y, Liu R, Xue W, Yang M, Shiraishi M, et al. Effective Fusion Method on Silhouette and Pose for Gait Recognition. IEEE Access. 2023;11:102623-102634.

[13] Castro FM, Marín-Jiménez MJ, Mata NG, Muñoz-Salinas R. Fisher Motion Descriptor for Multiview Gait Recognition. Int J Pattern Recognit Artif Intell. 2017;31:1756002.

[14] Fan C, Ma J, Jin D, Shen C, Yu S. Skeletongait: Gait Recognition Using Skeleton Maps. AAAI. 2024;38:1662-1669.

[15] Han J, Bhanu B. Individual Recognition Using Gait Energy Image. IEEE Trans Pattern Anal Mach Intell. 2006;28:316-322.

[16] Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y. Geinet: View-Invariant Gait Recognition Using a Convolutional Neural Network. In 2016 International Conference on Biometrics (ICB). IEEE. 2016:1-8.

[17] Zhang X, Xiong W, Yang Z, Zhang Q, Yan J. Gait Recognition by Combining Recurrent Neural Network and Fully Convolutional Network. Int J Pattern Recognit Artif Intell. 2025;39:2550006.

[18] Sun Y, Zheng L, Yang Y, Tian Q, Wang S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). 2018:480-496.

[19] Fu Y, Wei Y, Zhou Y, Shi H, Huang G, et al. Horizontal Pyramid Matching for Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:8295-8302.

[20] Luo H, Gu Y, Liao X, Lai S, Jiang W. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).IEEE/CVF. 2019:1487-1495,

[21] Qin H, Chen Z, Guo Q, Wu QJ, Lu M. Rpnet: Gait Recognition with Relationships between Each Body-Parts. IEEE Trans Circuits Syst Video Technol. 2021:2990-3000.

[22] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv preprint: https://arxiv.org/pdf/2010.11929

[23] Wu H, Tian J, Fu Y, Li B, Li X. Condition-Aware Comparison Scheme for Gait Recognition. IEEE Trans Image Process. 2021;30:2734-2744.

[24] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. Adv Neural Inf Process Syst. 2017;30.

[25] Jadon S. A Survey of Loss Functions for Semantic Segmentation. 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). IEEE. 2020:1-7.

[26] Schroff F, Kalenichenko D, Philbin J. Facenet: A Unified Embedding for Face Recognition and Clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE. 2015:815-823.

[27] Yu S, Tan D, Tan T. 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). IEEE. 2006;4:441-444.

[28] Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y. Multi-View Large Population Gait Dataset and Its Performance Evaluation for Cross-View Gait Recognition. IPSJ Trans Comput Vis Appl. 2018;10:1-14.

[29] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, et al. Pytorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neural Inf Process Syst. 2019;32.

[30] Li X, Makihara Y, Xu C, Yagi Y, Ren M. Joint Intensity Transformer Network for Gait Recognition Robust against Clothing and Carrying Status. IEEE Trans Inf Forensics Secur. 2019;14:3102-3115.

[31] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. 2014. ArXiv preprint: https://arxiv.org/pdf/1412.6980

[32] Zou S, Xiong J, Fan C, Shen C, Yu S, et al. A Multi-Stage Adaptive Feature Fusion Neural Network for Multimodal Gait Recognition. IEEE Trans Biom Behav Identity Sci. 2024;6:539-549.

[33] Zhu Z, Guo X, Yang T, Huang J, Deng J, et al. Gait Recognition in the Wild: A Benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE/CVF. 2021:14789–14799.

[34] Yu S, Huang Y, Wang L, Makihara Y, Reyes EB, et al. Competition on Human Identification at a Distance 2021. In 2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2021:1-7.