ORIGINAL RESEARCH PAPER

Physiotherapy

KEY WORDS: Gait, Step Length, Stride Length, Cadence, Gait Recognition, Gait Pattern.

GLGAIT: ENHANCING GAIT RECOGNITION WITH GLOBAL-LOCAL TEMPORAL RECEPTIVE FIELDS FOR IN-THE-WILD SCENARIOS

Anuj Kabra BPT, MPT (Orthopedics), Certified McKenzie Therapist, Keller Oaks Healthcare Services, Keller (Texas, USA)

ABSTRACT

Gait recognition has emerged as a promising non- intrusive human identification technology, capable of functioning from a distance without requiring active cooperation. While existing methods have shown considerable success in controlled laboratory environments, their performance often degrades in real-world scenarios. To address this challenge, we propose GLGait, a novel framework designed to improve gait recognition in the wild by establishing comprehensive temporal receptive fields. The core of GLGait is the Global-Local Temporal Module (GLTM), which integrates a Pseudo Global Temporal Self- Attention (PGTA) mechanism with temporal convolution operations. PGTA effectively captures pseudo-global temporal patterns with reduced computational overhead compared to traditional multi-head self-attention (MHSA), while temporal convolution enhances local temporal representations and aggregates them into a unified holistic temporal receptive field. Additionally, we introduce a Center-Augmented Triplet Loss (CTL), which minimizes intra-class distances and increases positive sample rep- resentation during training. Extensive evaluations on challenging in-the-wild datasets, such as Gait3D and GREW, demonstrate that GLGait achieves state-of-the-art performance, offering a robust and efficient solution for gait recognition in unconstrained settings. This work provides a significant step forward in making gait recognition systems reliable for real-world applications.

I. INTRODUCTION

Step acknowledgment, a biometric strategy for distinguishing people in view of their strolling designs, has earned critical consideration as of late. Not at all like other biometric identifiers, for example, facial elements or iris designs, walk Enjoys the extraordinary benefit of being perceivable from a good ways, without requiring the dynamic interest or consciousness of the subject. This ability makes walk acknowledgment especially engaging for applications where subtle observing is urgent. Throughout the long term, many high level appearance- based approaches have been created to Improve the precision of stride acknowledgment. These techniques fundamentally depend on investigating outline successions to separate special walk designs. Striking headways have been exhibited on controlled, in-the-lab datasets like CASIA-B and OU-MVLP. These datasets are portrayed by predictable natural circumstances, making it simpler for calculations to accomplish elevated degrees of execution. Be that as it may, the viability of these techniques lessens essentially when applied to genuine world, in-the-wild datasets like Gait3D and Developed. The significant drop in execution can be credited to the unmistakable difference between the controlled conditions of in-the-lab datasets and the erratic idea of in-the-wild situations. Not at all like the lab settings have true conditions presented a plenty of difficulties. People on foot might change their strolling speeds, digress from straight strolling ways, or experience impediments brought about by others or articles in the scene. These unusual and loud factors confound the errand of precisely catching and examining stride designs, prompting critical execution corruption in calculations fundamentally prepared on in-the-lab datasets. To resolve this issue, we started by leading a similar investigation of outline successions from the CASIA-B and Gait3D datasets. As outlined in, In-the-lab datasets like CASIA-B highlight different and equitably disseminated step cycles. These clear cut cycles give ideal transient fragments that empower calculations to actually learn and perceive total step designs. A neighborhood fleeting responsive field, which centers on a more modest, explicit fragment of the succession, frequently does the trick in such situations to catch the fundamental highlights of the walk cycle. In any case, in genuine situations, such perfectly characterized transient portions are seldom accessible. Varieties in strolling pace, shifts in course, and impediments fundamentally modify the presence of step designs, disturbing the consistency expected for customary techniques. These varieties require a more thorough way to deal with design acknowledgment. In particular, a worldwide fleeting open field becomes essential to adjust and examine step designs really. This worldwide viewpoint guarantees that even with abnormalities and commotion in the information, the calculation can catch the general construction of a singular's stride, empowering powerful acknowledgment even in testing conditions. Recently, several methods [10], [11] have been proposed to address the issue of gait recognition in the wild. These methods primarily leverage Convolutional Neural Networks (ConvNets).

Compared to methods designed for in-the-lab datasets, they often incorporate more convolutional operations to exhibit a larger receptive field. Given that gait silhouettes are typically small (e.g., 64×64 pixels), the spatial receptive field of these methods is generally sufficient. However, the temporal receptive field obtained through convolutional operations is significantly limited. As shown in Figure 1(b), the temporal receptive fields can usually only cover approximately 50 silhouettes in a sequence, which is insufficient given the typical frame numbers in in-the-wild gait sequences. The limited local temporal receptive field.

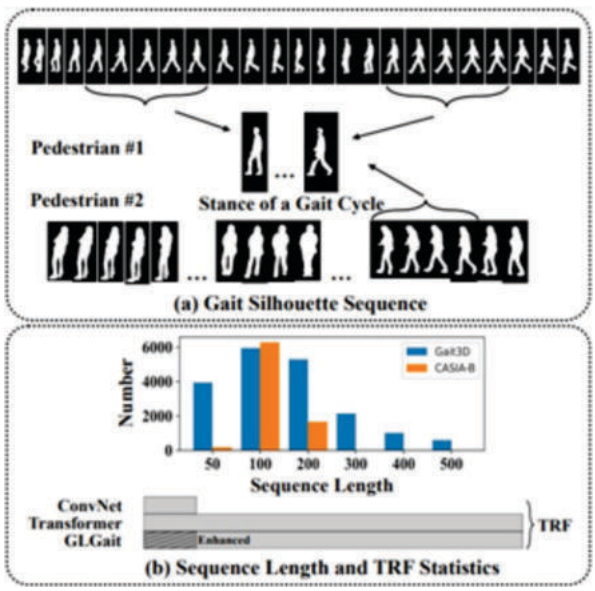

Fig. 1. Comparison of local and global temporal receptive field. (a) Gait cycles are evenly distributed in laboratory scenarios (Pedestrian #1), thus proper-sized local receptive field can capture a complete cycle. While in the wild (Pedestrian #2) the distribution is sparse and random, which

implies a larger receptive field is necessary. Corresponding sequences are sampled from CASIA-B [51] and Gait3D [52], respectively. (b) Sequence length and TRF statistics in CASIA-B and Gait3D, where TRF is the temporal receptive field.

Fails to capture adequate information about pedestrian body shape changes, necessitating a global temporal receptive field. Some works [2] utilize visual transformer blocks [7] to obtain a global temporal receptive field. However, directly replacing convolution blocks with transformer blocks does not necessarily enhance the local temporal receptive field. While transformer blocks use multi-head self-attention (MHSA) to obtain a global receptive field, adding MHSA before temporal convolution operations in ConvNets could combine global and local temporal receptive fields. Nevertheless, using the output of ConvNets as input to MHSA results in dimension explosion due to large channels (e.g., 512), leading to high memory consumption and computational cost. To address this issue, we propose Pseudo Global Temporal Self-Attention (PGTA). Compared to MHSA, PGTA reduces complexity in two ways. First, to address the temporal receptive field issue, we separate the spatial dimension, focusing only on the temporal patch size. Second, inspired by, we separate the patch size from tokens in PGTA, creating a pseudo temporal receptive field for each token. These pseudo global temporal receptive fields naturally aggregate into a true global temporal receptive field with the same temporal convolution kernel size and patch size. We combine PGTA with temporal convolution operations in a module called the Global-Local Temporal Module (GLTM). Based on GLTM, we design a Global-Local Temporal Receptive Field Network, named GLGait. The backbone of GLGait consists of a vision encoder and Global-Local 3D (GL-3D) blocks. In the vision encoder, GLGait encodes a preliminary pedestrian representation. For the GL-3D blocks, GLTM is used in the temporal dimension to effectively obtain a global-local temporal receptive field, while 2D convolution operations are applied in the spatial dimension since the spatial receptive field is already sufficient. Furthermore, inspired by center loss [17], we propose a simple yet effective loss function, named Center-Augmented Triplet Loss (CTL), to assist in model training. CTL extends conventional triplet loss [18] by incorporating the class center as a positive sample for each input. This approach has two ad- vantages: reducing intra-class distance and expanding positive samples during training.

Contributions. The main contributions of this work are summarized as follows:

1) We design a Global-Local Temporal Receptive Field Network (GLGait) to obtain a global-local temporal receptive field for gait recognition in the wild.
2) We propose Pseudo Global Temporal Self-Attention (PGTA) to reduce the high memory and computational complexity of multi-head self-attention (MHSA) .
3) We introduce a Center-Augmented Triplet Loss (CTL) to assist in model training by reducing intra-class distance and expanding positive samples.
4) Extensive experiments demonstrate that our approach achieves state-of-the-art performance on in-the-wild datasets, e.g., Gait3D and GREW.

II. Related Works

A. Model-based Stride Recognition

Model-based techniques [1, 13, 25-27, 35, 41, 42, 49] in walk acknowledgment center around figuring out the actual design of the human body, utilizing itemized present data to upgrade stride examination. These techniques normally depend on portrayals like 2D skeletons, 3D joint information, or even point mists got from different sensor modalities. By taking into account the physical elements of the human body and its developments, model-based strategies can give more precise and hearty walk acknowledgment, particularly in unique or complex conditions. For instance, PoseGait [27] uses human body present data to separate worldly spatial elements, which are then handled utilizing Convolutional Brain Organizations (ConvNets) to catch significant level highlights from the arrangement of body presents. This approach features the significance of perceiving the spatial connection between various body parts after some time, which can be especially important for recognizing stride designs. Likewise, GaitGraph [42] presents a posture assessor that concentrates present elements from the information and utilizes diagram convolutional networks (GCNs) to break down the stride. GCNs are adroit at learning the connections between various joints or body portions in the human posture, consequently working on the model's capacity to catch both neighborhood and worldwide conditions in walk acknowledgment.

Another striking methodology is GPGait [13], which uses a bound together posture portrayal as contribution to the model. This strategy presents a section mindful chart convolutional network (PAGCN) to proficiently parcel the diagram into more modest districts for better neighborhood highlight extraction, while likewise saving worldwide spatial elements. This double spotlight on nearby and worldwide spatial data makes GPGait especially viable in dealing with complex human movement information. LiDAR-based techniques have additionally been investigated in walk acknowledgment. For instance, LidarGait [38] use LiDAR sensor information to produce point mists that address the step of a person. This approach benefits from the exact 3D information given by LiDAR, which helps in catching nitty gritty walk designs even in testing conditions like changing light levels or impediments. One of the critical benefits of model-based techniques is their heartiness in situations where outside factors, like changes in attire, could affect appearance-based strategies. Studies [5, 37] have shown the way that model-based approaches can in any case precisely perceive walk in any event, when the subject changes clothing, as they center more around the development examples and physical design as opposed to visual highlights that could be clouded by clothing changes. This heartiness makes them a promising arrangement in genuine applications where such varieties are normal. Not with standing, in spite of their assets, these techniques face specific difficulties. Present assessment, a urgent move toward model-based stride acknowledgment, can be computationally concentrated and may require specific hardware or calculations for exact following. The intricacy of ascertaining and deciphering present data can be a deterrent in certain settings, particularly in new or dynamic conditions where present assessment models may not be prepared for explicit body types or developments. Thus, applying model-based step acknowledgment strategies to completely new situations, for example, novel strolling surfaces or different natural circumstances, can introduce hardships. Moreover, these models frequently require significant computational assets and are subject to the nature of info information, making them less adaptable than appearance-based approaches in specific ongoing applications.

B. Appearance-based Stride Recognition

Appearance-based techniques in walk acknowledgment expect to learn and examine the body's appearance without requiring express primary data. These techniques center around extricating highlights from the visual appearance of a subject's walk, regularly utilizing outline groupings or other obvious signals got from video or picture information. The benefit of appearance-based strategies lies in their effortlessness and productivity, as they don't need nitty gritty posture data or complex physical demonstrating. Rather, these strategies depend on profound brain organizations to gain discriminative elements from the info outline arrangements, considering step acknowledgment by contrasting these highlights and others from various walk groupings.

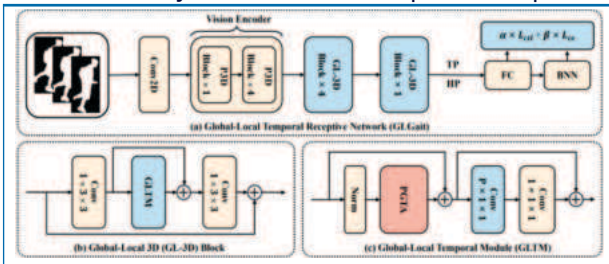

Fig. 2. Pipeline of the proposed GLGait. The backbone mainly consists of the vision encoder and GL-3D blocks. Specifically, we use Pseudo Global Temporal Self-Attention (PGTA) to extract global temporal information and a temporal convolution operation to enhance the local temporal information extraction in Global-Local Temporal Module (GLTM). TP denotes the Tem- poral Max Pooling operation, HP is the Horizontal Pooling operation [11, 14], FC is the separate fully connected layers [4], and BNN is BNNeck [32]. The final loss function is composed of a center-augmented triplet loss (CTL) and a cross-entropy loss.

Numerous appearance-based techniques depend on outline arrangements as the essential info. These arrangements are produced from the body's diagrams as caught through different imaging frameworks, like RGB cameras or profundity sensors. By handling these groupings utilizing profound learning models, for example, convolutional brain organizations (CNNs), analysts can naturally extricate significant step highlights. These elements are then contrasted with those from other stride successions to perform acknowledgment. Along these lines, appearance put together techniques center with respect to the visual examples of strolling, utilizing the peculiarity of a singular's step caught in outline structure.

Besides, a few techniques have progressed the field by integrating semantic parsing of walkers into their models. These methods dissect the fundamental semantics of the person on foot's appearance, removing extra highlights that assistance to further develop acknowledgment precision. These models think about more elevated level obvious signals, for example, body part division, clothing style, or even movement designs, offering more strong execution in complex certifiable situations. By and large, most appearance-based walk acknowledgment procedures have been assessed in controlled, lab conditions. Early works zeroed in on these in-the-lab situations, where subjects were much of the time caught in profoundly controlled conditions, prompting high precision in acknowledgment assignments. Notwithstanding, with the developing accessibility of more different and complex genuine world datasets, for example, Gait3D and Developed, appearance-based techniques are confronting new difficulties. These"in nature" datasets present greater fluctuation in common appearance, ecological factors, and strolling conditions, which confuses stride acknowledgment errands. To address these difficulties, specialists have proposed a few techniques to work on the vigor and exactness of appearance- based stride acknowledgment in additional dynamic and differed conditions. For example, GaitGCI acquaints counterfactual mediation learning with relieve the effect of confounders, for example, foundation commotion or natural interruptions, while zeroing in on learning discriminative highlights in the most educational locales of the outline arrangement. This assists with making the model more interpretable and precise in recognizing key step qualities. Essentially, DyGait centers around unique elements, fostering a powerful expansion module that advances part-explicit highlights naturally, in this way working on the model's capacity to deal with varieties in strolling styles or ecological changes. Another striking commitment is from Fan et al. [10], who propose a basic yet successful ResNet-like structure, GaitBase, which use remaining figuring out how to further develop acknowledgment precision. Expanding on this, DGaitV2 [10] develops the first GaitBase model by stacking extra blocks, further improving its presentation and empowering it to accomplish higher exactness on more testing datasets. These advances have permitted appearance-based techniques to expand their relevance past controlled settings, making them more versatile to genuine situations. One of the fundamental benefits of appearance-based techniques, particularly when contrasted with model-based strategies, is their simplicity of organization in new situations. Since appearance highlights are gotten straightforwardly from visual information, they are somewhat simple to catch utilizing normal imaging frameworks, like standard cameras or profundity sensors. This gives more extensive versatility and adaptability, permitting these strategies to be utilized across various conditions and for various sorts of people on foot. Besides, appearance-based techniques normally require less computational power contrasted with model-based strategies, making them more commonsense for continuous applications. Notwithstanding, regardless of these advantages, appearance-based techniques are as yet defenseless to specific restrictions. One significant test is their dependence on the visual appearance of the walker, which can be handily impacted by variables like changes in apparel, conveying packs, or other individual assets. In unambiguous situations, for example, a subject changing their dress appearance-based strategies can experience the ill effects of decreased power, as the step highlights extricated from outlines might be essentially modified by these visual changes. This weakness to changes in appearance features the compromise among versatility and vigor, which is a typical worry in numerous PC vision applications. Our own methodology, GLGait, has a place with the appearance-based classification, utilizing step outline successions as contribution for acknowledgment. GLGait presents an imaginative technique by laying out a worldwide nearby transient responsive field, which catches both the general step qualities and fine-grained subtleties of neighborhood movement designs. This double level center improves the model's capacity to separate between people while keeping up with hearty execution across different circumstances. By joining worldwide and neighborhood fleeting elements, GL- Gait expects to address a portion of the difficulties looked by customary appearance-based strategies, offering a more thorough answer for walk acknowledgment in both controlled and true conditions.

C. Gait Transformers

Vision transformer [7] has achieved successful performance in many fields, such as classification [7], objective detection [29], and semantic segmentation [21]. Some works introduce transformer blocks into the gait recognition framework. Trans- Gait [24] proposes a set transformer model with a temporal aggregation operation for obtaining set-level spatio-temporal features. SwinGait [10] utilizes convolutional blocks to extract silhouette feature and feed it to swinformer [30, 31] blocks. However, the former cannot enhance the local temporal receptive field, while the latter only obtains a window-global temporal receptive field. Differently, our GLGait employs Global-Local Temporal Module (GLTM) to both maintain global-local temporal receptive fields.

III. Method

In this section, we first introduce the pipeline of GLGait. Then, we detail the vision encoder, Global-Local 3D (GL-3D) block, and centeraugmented triplet loss (CTL), respectively. Finally, we explain the optimization.

A. Pipeline

The pipeline of the proposed GLGait is shown in Figure 2 (a). A simple 2D convolution operation first initializes the silhouette sequences. Then we utilize the vision encoder to obtain a preliminary representation. After that, GL-3D blocks are used to extract both spatial information and global-local

temporal information. Temporal Max Pooling operation (TP) and Horizontal Pooling operation [11, 14] are employed to aggregate the features. Finally, a combined loss function consisting of center-augmented triplet loss and cross-entropy loss is used to supervise the learning process.

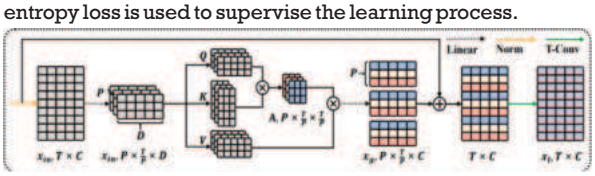

Fig. 3. Pseudo Global Temporal Self-Attention (PGTA) with a Temporal Convolution Operation (T-Conv).

B. Vision Encoder

Since gait silhouettes are binary and contain limited information [10], we utilize a vision encoder to encode a preliminary pedestrian representation. Specifically, we employ the conventional Pseudo 3D (P3D) blocks [16] as the primary components. The P3D block uses two 2D convolutions in the spatial dimension and a 1D convolution in the temporal dimension. Given an input $x_{in} \in R^{C\times T\times H\times W}$, where C represents the number of channels, T is the number of frames, and H and W denote the height and width of the silhouette, respectively, the P3D block processes the spatial and temporal dimensions separately. This separation ensures efficient representation learning while preserving the unique characteristics of the gait silhouette.

C. GL-3D Block

Operation: As shown in Figure 2(b), the GL-3D block consists of two 2D convolutions for spatial information and a Global-Local Temporal Module (GLTM) for temporal features. Specifically, GLTM mainly contains Pseudo Global Temporal Self-Attention (PGTA) and a temporal convolution operation, as shown in Figure 2(c). Unlike normal multi-head self-attention (MHSA) , PGTA focuses on the temporal dimension, reducing memory and computational complexity in two aspects:

- PGTA separates the spatial dimension from the patch size, making it 1D instead of 3D.
- PGTA separates the patch size from the token, inspired by.

The structure of PGTA is shown in Figure 3, where only one head is considered for simplicity.

Given an input $\mathbf{x}_{in} \in \mathbf{R}^{L\times T\times C}$, where $L = H\times W$, the formula for PGTA is as follow:

$$\text{Reshaping: } \mathbf{x}_{in}, \mathbf{R}^{L\times T\times C} \rightarrow \mathbf{R}^{L\times P\times \frac{T}{P}\times C} \quad (1)$$

$$[\mathbf{q}_i, \mathbf{k}_i, \mathbf{v}_i] = \mathbf{x}_{in}, \mathbf{U}_{qkv}, \ \mathbf{U}_{qkv} \in \mathbf{R}_{C\times 3D} \quad (2)$$

$$\mathbf{A}_i = \text{softmax}\ \frac{\sqrt{\mathbf{q}_i\ \mathbf{k}^{\mathrm{T}}}}{\mathrm{D}}, \ \mathbf{A}_i \in \mathbf{R}^{L\times P\times \frac{T}{P}\times \frac{C}{P}} \quad (3)$$

$$\mathbf{X}_i = \mathbf{A}_i\ \mathbf{v}_i, \ \mathbf{X}_i \in \mathbf{R}^{L\times P\times \frac{T}{P}\times D}, \quad (4)$$

$$\mathbf{x}_g = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_k]\ \mathbf{U}_{msa}, \ \mathbf{U}_{msa} \in \mathbf{R}^{kD\times C} \quad (5)$$

$$\text{Reshaping: } \mathbf{x}_g\ \mathbf{R}^{L\times P\times \frac{T}{P}\times C} \rightarrow \mathbf{R}^{L\times T\times C} \quad (6)$$

where $i \in \{1, 2, \ldots, k\}$, k is the number of heads in PGTA, P is the patch size, A_i is the attention matrix and U_{qkv} and U_{msa} are parameter matrices. After PGTA, a temporal convolution operation enhances the local temporal receptive field. Inspired by , we add skip connections and a 1D convolution operation with a kernel size of one as a linear layer after PGTA to compute a sub-update for each token element. This process is formulated as:

$$\mathbf{x}_l = R(\mathrm{C1D}(\mathbf{x}_g + \mathbf{x}_{\mathrm{in}})), \ \mathbf{x}_l \in \mathbf{R}^{L\times T\times D_{\mathrm{mid}}} \quad (7)$$

$$\mathbf{x}_{out} = \mathbf{x}_l \mathbf{U}_{mlp} + \mathbf{x}_g, \ \mathbf{U}_{mlp} \quad \mathbf{R}^{D_{\mathrm{mid}}\times C}, \quad (8)$$

where $x_{out} \in R^{L\times T\times C}$ is the output of GLTM, R and C1D are the same as operations in Equation (1) and Equation (6), D_{mid} is the number of hidden channels, and U_{mlp} is a parameter matrix. We analyze the memory and computational complexity of PGTA compared to normal Spatio-Temporal MHSA [2]. Specifically, we consider two cases for D in Equation (2):

- D is equal to the token size.
- D is a scalar smaller than C.

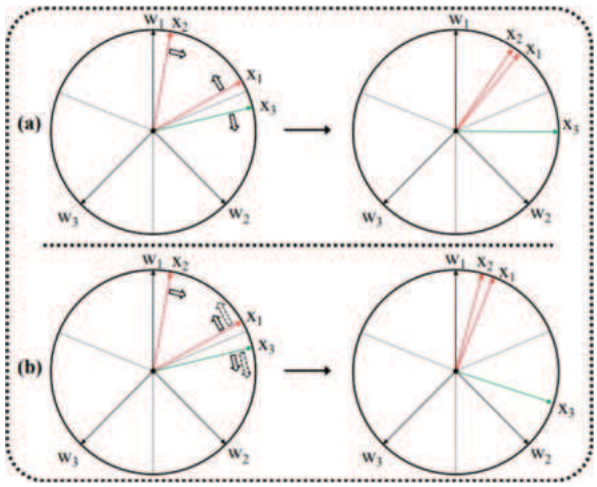

Fig. 4. Comparison of Triplet Loss

Given $\mathbf{x}_{in} \in \mathbf{R}^{L\times T\times C}$ and token size $P\times C$ the computation complexity for MHSA and PGTA are as follows:

$$C_{MHSA}^{mem} = O(P^2C^2) \quad (9)$$

$$C_{MHSA}^{comp} = O(L\ T\ C), \quad (10)$$

$$C_{PGTA}^{mem} = O(PC^2), \quad (11)$$

$$C_{PGTA}^{comp} = O(LT^2C). \quad (12)$$

By setting $P_t = 3$ and $P_l = 4$ in the experiment, we observe that PGTA reduces memory complexity from O ($P_l^2P_t^2C^2$) to O (P_tC^2). Similarly, computation complexity is reduced to $^1/_{16}$ of MHSA. This demonstrates the efficiency of PGTA, especially when D is smaller than C, as it further reduces complexity. For MHSA, given the token size $P_l\times P_t\times C$, the memory complexity is O (P_lP_tCD) and computational complexity is O($L^2T^2/P_l^2P_t^2D$). For PGTA, given the token size $P_t\times C$ the memory complexity is O(CD) and computational complexity is O($LT^2/\ P_tD$). PGTA significantly reduces the memory complexity from O(P_lP_tCD) to O (CD). For computational complexity, the reduction is influenced by L and $P_l\times P_t$. Using the same setting as above, PGTA reduces the computational complexity to one quarter of MHSA. Furthermore, MHSA suffers from considerable information loss. Assuming the feature dimension serves as the measure, the information loss of PGTA is only O(C-D). In contrast, for MHSA, the information loss reaches O(P_lP_tC-D).

Finally, regarding memory complexity, we set D as a scalar in all our experiments to minimize overhead.

D. Center-Augmented Triplet Loss

Operation: In contrast to the conventional triplet loss [18], the center-augmented triplet loss (CTL) incorporates class centers as positive instances for each sample. The formulation is given as:

$$L_{etl}(\mathrm{x}) = \sum_{i=1}^{\succeq} \sum_{j=1}^{\succeq} \max(D(x, x_q) - D(x, x_n) + m, 0), \quad (13)$$

$$L_{\mathrm{ctl}} = \frac{1}{B} \sum_{k=1}^{\succeq B} L_{\mathrm{ctl}}(x_k), \quad (14)$$

where $x \in R^C$ is the input, x_q is the i-th positive sample of x, x_{nj} is the j-th negative sample of x, D(,) is the distance function, m is the margin, Q is the number of positive samples, N is the number of negative samples, and B is the batch size (B=Q+N +1). Specifically, x_{qQ+1} is the class center of x, and Euclidean distance is used for D(., .).

DISCUSSION

In this subsection, we discuss the mechanism of the center-augmented triplet loss (CTL). Beyond the common effects of triplet loss [18], CTL offers two advantages:

First, CTL reduces intra-class distance. As illustrated in Figure

4(a), for sample x_1, x_2 is a positive sample, and x_3 is a negative one. When the distance between (x_1, x_2) exceeds that of (x_1, x_3), the triplet loss computes the corresponding gradient (represented as solid-line arrows), encouraging (x_1, x_2) to move closer and (x_1, x_3) to move apart. However, as x_2 moves closer to x_1, it moves further away from the class center w_1, thus increasing the intra-class distance. Conversely, as shown in Figure 4(b), by treating w_1 as a positive sample for x_1 and w_3 as one for x_3, CTL calculates respective gradients (depicted as dashed-line arrows), driving x_1 towards w_1 and x_3 towards w_3. Due to the combined gradients from (x_1, x_2) and (x_1, w_1), x_1 moves closer to w_1, indirectly preventing x_2 from moving away from the class center w_1, thus reducing the intra-class distance.

Secondly, CTL directly increases the number of positive samples without expanding the batch size. In gait datasets such as Gait3D [52], many pedestrian silhouette sequences are limited in quantity, leading to a lack of positive samples in a batch. CTL utilizes class centers as positive samples, adding computational complexity to the loss function without affecting the backbone computations during model training or model inference during testing. This presents a cost-effective trade-off.

E. Optimization

Training Stage. In the training stage of GLGait, a combined loss function consisting of the center augmented triplet loss (L_{ctl}) and cross-entropy loss (L_{ce}) is calculated to supervise the learning process:

$$L = \alpha L_{ctl} + \beta L_{ce}, (15)$$

where α and β are hyper parameters to balance the contributions to the total loss L.

IV. Experiments

In this section, we first introduce the datasets and implementation details. Then, we compare our proposed GLGait with the latest gait recognition methods and analyze the results. Finally, extensive ablation experiments prove the effectiveness of each component in GLGait.

A. Datasets and Implementation Details

The dataset information and implementation details in our experiments are as follows. Gait3D [52] is a large scale gait dataset. Within a supermarket, 39 cameras capture 1,090 hours of videos with 1,920×1,080 resolution and 25 FPS. Through processing, a total of 4,000 subjects, 25,309 sequences, and 3,279,239 frame images are extracted. 3,000 subjects are compiled as the training set, while the remaining 1,000 subjects form the test set. For the testing phase, the probe comprises one sequence from each subject, and the gallery consists of the rest sequences. GREW [54] is one of the largest gait datasets in the wild, including Silhouettes, GEIs, and 2D/3D human poses data types. The raw videos are collected from 882 cameras in large public areas. 7,533 video clips are used, containing nearly 3,500 hours of 1,920×1,080 streams. It has 26,345 subjects and 128,671 sequences, divided into two parts with 20,000 and 6,000 subjects as training set and test set,

B. Datasets and Execution Details

In this part, we give thorough data with respect to the datasets utilized in our examinations, as well as the execution subtleties of the models and preparing conventions.

Gait3D Dataset: The Gait3D dataset is a huge scope dataset explicitly intended for walk acknowledgment. It is gathered in a genuine grocery store setting, where a sum of 39 cameras catch roughly 1,090 hours of video film. The video goal is 1,920×1,080 with a casing pace of 25 edges each second (FPS). The dataset comprises of more than 4,000 subjects, with a sum of 25,309 stride successions and 3,279,239 individual edges separated. For the preparation stage, we utilize 3,000 subjects, while the leftover 1,000 subjects are saved for testing. In the testing stage, the test set comprises of one grouping from each subject, and the exhibition set incorporates all excess successions of those subjects. This dataset is intended to give a different scope of step groupings for both preparation and testing, guaranteeing strong assessment across different situations.

GREW Dataset: The Developed dataset is one of the biggest stride acknowledgment datasets gathered in uncontrolled, true conditions. The crude video cuts in this dataset are caught from 882 cameras dispersed across huge public spaces, adding an elevated degree of variety and intricacy to the information. Developed incorporates various information types, including outline pictures, Walk Energy Pictures (GEIs), and 2D/3D human posture information. The dataset comprises of 7,533 video cuts that together range almost 3,500 hours of 1,920×1,080 goal video transfers. It contains 26,345 exceptional subjects, with 128,671 step arrangements altogether. The dataset is separated into two sections: a preparation set with 20,000 subjects and a test set with 6,000 subjects. This dispersion guarantees that the models are prepared on a different scope of subjects while as yet giving adequate information to thorough testing and assessment. Given the uncontrolled idea of the dataset, Developed presents huge difficulties for stride acknowledgment, making it an optimal benchmark for assessing the speculation capacity of acknowledgment models in certifiable situations.

Implementation Details: Our tests were directed utilizing the PyTorch structure, utilizing its useful assets for profound learning and model turn of events. The organization engineering was planned in light of earlier work in stride acknowledgment [10], [11], and we followed their pattern models with slight adjustments, as displayed in Table I . The boundary values utilized in the tests were set as follows: in Condition (22), both □ and □ were set to 1. The bit size for all convolution tasks was set to 3, which is a standard decision for guaranteeing productive component extraction while keeping up with computational possibility. In Condition (8), the boundary P was set to 3, and D was arranged to match the quantity of channels C in Stage-1. Additionally, in Condition (11), D_{mid} was likewise set to a similar worth as C in all stages to guarantee consistency in highlight portrayal across the organization. For boundary determination, we zeroed in just on the foundation of the model, barring the completely associated (FC) layers [4] and Bayesian Brain Organization (BNN) layers in all analyses. This approach permits us to examine the center exhibition of the model without presenting extra intricacies that could darken the fundamental way of behaving of the design. Like DGaitV2 [10], we parcel the model limit into three unmistakable sections to figure out some kind of harmony between acknowledgment exactness and computational expense. These sections are GLGait-Base (GLGait-B), GLGait-Enormous (GLGait-L), and GLGait-Colossal (GLGait-H). Every one of these models has a similar design, with the main contrast being the quantity of channels utilized at each stage. The channel designs for these models are as per the following: GLGait-B has (32, 64, 128, 256), GLGait-L has (64, 128, 256,512), and GLGait-H has (128, 256, 512, 1024). This variety in channel numbers permits us to control the limit of the models and analysis with various compromises among precision and computational necessities. For all preparing tests, the info outlines were resized to a decent size of 64 44. Moreover, each outline grouping was requested with a length of 30 edges. This reliable grouping length guarantees that the models are prepared on a uniform info size and can learn worldly conditions across the stride successions.

Optimizer and Preparing Protocol: We used the Stochastic Angle Plummet (SGD) analyzer to prepare the models. The SGD enhancer was picked for its unwavering quality and viability in huge scope profound learning assignments.

Table I Comparisons Of Rank-1, Rank-5, And Rank-10 Accuracies, Mean Average Precision (MAP) (%), And Parameter Size (M) On Gait3d And Grew Datasets.

Backbone	Method	Source	Gait3D Rank-1	Rank-5	mAP	GREW Rank-1	Rank-5	Rank-10	Params (M)
Convolution	GaitSet [4]	AAAI 2019	36.7	58.3	30.0	46.3	63.6	70.3	2.56
	GaitPart [12]	CVPR 2020	28.2	47.6	21.6	44.0	60.7	67.3	1.46
	GaitGL	ICCV 2021	29.7	48.5	22.3	47.3	63.6	69.3	2.49
	SMPLGait	CVPR 2022	42.9	63.9	35.2	-	-	-	-
	DyGait	ICCV 2023	66.3	80.8	56.4	71.4	83.2	86.8	-
	HSTL	ICCV 2023	61.3	76.3	55.5	62.7	76.6	81.3	4.05
	GaitGCI [9]	CVPR 2023	50.3	68.5	39.5	68.5	80.8	84.9	-
	GaitBase [11]	CVPR 2023	64.3	79.6	55.5	59.1	74.5	78.9	4.90
Convolution	DGaitV2-2D-B [10]	Arxiv 2023	64.5	81.7	56.5	62.3	76.4	81.5	2.35
	DGaitV2-2D-L [10]	-	67.8	83.9	59.7	69.7	82.4	86.7	9.33
	DGaitV2-P3D-B [10]	-	70.8	85.7	62.9	72.6	84.5	87.9	2.79
	DGaitV2-P3D-L [10]	-	74.2	86.9	67.1	78.3	88.5	91.4	11.12
	DGaitV2-P3D-H [10]	-	75.0	-	-	81.0	-	-	44.43
	DGaitV2-3D-B [10]	-	71.0	85.0	62.3	73.1	84.9	88.4	6.92
	DGaitV2-3D-L [10]	-	74.1	87.0	66.5	79.0	88.9	91.6	27.62
	DGaitV2-3D-H [10]	-	75.8	-	-	81.6	-	-	110.44
Convolution +	SwinGait-3D [10]	Arxiv 2023	75.0	86.7	67.2	79.3	88.9	91.8	13.10
	TrGaLnsGfoaritm-Ber	Ours	73.9	86.3	65.9	75.4	86.3	89.6	3.58
	GLGait-L	-	77.6	88.4	69.6	80.0	89.4	92.2	14.28
	GLGait-H	-	77.7	88.9	70.6	82.8	91.1	93.5	57.04

We applied a weight rot of 0.0005 and an energy of 0.9 to further develop intermingling and forestall over fitting. During preparing, we changed the learning rate, clump size, and number of emphases to oblige the differing sizes of the datasets:

Gait3D: For the Gait3D dataset, the model was prepared for a sum of 120,000 cycles, with a group size of 32 4 (32 people on foot, each containing 4 successions). The learning rate was introduced at 0.1 and diminished by a variable of 0.1 at emphases 40k, 80k, and 100k to permit the model to combine all the more really in the later phases of preparing. **GREW:** For the Developed dataset, the model was prepared for 180,000 cycles with a bunch size of 32 4. The learning rate was at first set to 0.05 and diminished by an element of 0.2 at cycles 60k, 120k, and 150k, following a comparative learning rate plan as Gait3D to keep up with viable combination all through the preparation interaction.

We likewise carried out and prepared the models from GaitBase [11] and DGaitV2 [10] ourselves to guarantee reliable assessment across various methodologies. The outcomes acquired in our tests are displayed in Table III, where the numbers on the left half of the brackets address our outcomes, and the qualities inside the enclosures relate to those detailed in the first papers [10], [11]. This correlation considers an unmistakable comprehension of how our models perform comparative with existing cutting edge draws near.

C. Performance Comparison

Methods Using Silhouette Sequence as Input on In- the-Wild Datasets: These methods, as shown in Table I, utilize silhouette sequences as inputs. Specifically, the input to SMPLGait exclusively comprises silhouettes. These methods can be divided into two categories based on their network backbone components: (1) backbones predominantly consisting of convolutional operations and (2) backbones constituted of both convolutional operations and transformers. For the first category, the receptive field plays an essential role. Regarding the spatial receptive field, as illustrated in Table I, despite DGaitV2-2D-B [10] having half the parameters of GaitBase [11], its spatial receptive field substantially exceeds that of GaitBase. Consequently, DGaitV2-2D- B surpasses GaitBase in accuracy on Gait3D and GREW by 0.2% and 3.2%, respectively. The second consideration is the temporal receptive field. Under equivalent spatial receptive field conditions, DGaitV2-P3D-B outperforms DGaitV2-2D-B by 6.3% and 10.3% in Rank-1 accuracy on Gait3D and GREW, respectively, with only a 0.44 MB parameter increase. A similar trend is observed with DGaitV2-P3D-L and DGaitV2- 2D-L. However, whether for DGaitV2-P3D or DGaitV2-3D, their temporal receptive fields are significantly insufficient compared to sequences extending hundreds of silhouettes, thereby extracting only limited local temporal information.

For the second category, the global-local temporal receptive field is of paramount importance. As depicted in Figure 1(a), gait exhibits a cyclical pattern, where each cycle represents a local silhouette sequence within the gait sequence. SwinGait-3D [10] utilizes a 3D residual block to encode a preliminary pedestrian representation, which is then fed into a 3D Swin former block, achieving a window-global temporal receptive field. However, within these blocks, SwinGait-3D cannot obtain a true global one. In contrast, GLGait utilizes GLTM to exhibit a global-local temporal receptive field, thereby more effectively learning the cyclical motions of gait. With similar parameter counts, GLGait surpasses SwinGait-3D in Rank- 1 accuracy both on Gait3D and GREW by 1.6% and 0.4%. Moreover, we conduct an extended experiment to further explore the performance of GLGait across varying sequence lengths. The results are shown in Table 3, where the length distribution is illustrated in Figure 1 (b). GLGait-L outperforms DGaitV2-P3D-L 5.1% and 10.7% Rank-1 accuracy at lengths 301 to 400 and 401 to 500, respectively. This demonstrates that GLGait is effective in long sequences. Besides, we also observe that GLGait improves 2.9% Rank-1 accuracy at lengths 1 to 100, indicating that GLGait is even effective in short sequences rather than only in long sequences. Finally, with the incorporation of CTL, GLGait-H achieves state of-the- art performance on both Gait3D and GREW, obtaining Rank1 accuracy of 77.7% and 82.8%, respectively. Effectiveness of Center-Augmented Triplet Loss. To verify the effectiveness of CTL, we conduct ablation experiments on DGaitV2-P3D [10] and GLGait. As shown in Table 4, CTL improves both DGaitV2-P3D and GLGait compared with conventional triplet loss [18] on Gait3D [52] and GREW [54], demonstrating its generalizability and effectiveness. Meanwhile, we also compare CTL with center loss [48] (CL) and triplet center loss [17] (TCL) as shown in Table 5. CTL outperforms them in GLGait- B and GLGait-L. The reason lies in that CL and TCL only focus on the connection between samples and class centers, ignoring the pair of samples to samples. In contrast, CTL considers the pair of both samples to samples and samples to class centers, reducing intra-class distance and expanding positive samples. CTL can seamlessly substitute conventional triplet loss [18], and we substitute it with CTL in subsequent experiments.

Table II Network Backbone Of GLGait.

Layer Name	Output Size	Structure
Conv2D	(T, C, 64, 44)	[1 × 3 × 3, C] × 1
Stage-1	Vision Encoder	(T, C, 64, 44)
		[1 × 3 × 3, C]
		[3 × 1 × 1, C]
		[1 × 3 × 3, C] × 1
Stage-2	Vision Encoder	(T, 2C, 32, 22)
		[1 × 3 × 3, 2C]
		[3 × 1 × 1, 2C]
		[1 × 3 × 3, 2C] × 4
Stage-3	GL-3D Block	(T, 4C, 16, 11)
		[1 × 3 × 3, 4C]
		[GLTM, 4C]
		[1 × 3 × 3, 4C] × 4
Stage-4	GL-3D Block	(T, 8C, 16, 11)
		[1 × 3 × 3, 8C]
		[GLTM, 8C]
		[1 × 3 × 3, 8C] × 1

Table III Extended Experiment Of Sequence Length On Gait3dWith Rank-1 Accuracy (%).

Method	Sequence Length	Rank-1 Accuracy (%)
	1-100	68.8
	101-200	84.1
	201-300	75.3
	301-400	83.0
	401-500	75.0
	1-500	74.2
DGaitV2-P3D-L		
	1-100	71.7
	101-200	84.1
	201-300	76.4
	301-400	88.1
	401-500	85.7
	1-500	76.6
GLGait-L		

Table IV Performance Gain From Applying CTL On Gait3d And GrewWith Rank-1 Accuracy (%).

Method	Gait3D	GREW
DGaitV2-P3D-B [10]	70.8 → 72.1	72.6 → 74.3
DGaitV2-P3D-L [10]	74.2 → 75.4	78.3 → 79.6
GLGait-B	73.3 → 73.9	74.2 → 75.4
GLGait-L	76.6 → 77.6	79.7 → 80.0

Table V Compared Center-augmented Triplet Loss (CTl) With Other Loss Function On Gait3d With Rank-1 Accuracy (%), Where Tl Is Triplet Loss [18], CL Is Center Loss, Tcl Is Triplet Center Loss [17].

Method	TL	CL	TCL	CTL	Rank-1
GLGait-B	✓	✗	✗	✗	73.3
GLGait-B	✗	✓	✗	✗	72.8
GLGait-B	✗	✗	✓	✗	73.3
GLGait-B	✗	✗	✗	✓	73.9
GLGait-L	✓	✗	✗	✗	76.6
GLGait-L	✗	✓	✗	✗	76.1
GLGait-L	✗	✗	✓	✗	76.2
GLGait-L	✗	✗	✗	✓	77.6

D. Ablation Experiments

We exhibit ablation experiments in GLGait to prove the effectiveness of each component. Vision Encoder Size. To explore an appropriate vision encoder size, we conduct ablation studies within a controlled network, where the number of channels and blocks in per stage are fixed. Specifically, we employed the P3D block [16, 36] as the component of the vision encoder. As shown in Table 6, the model demonstrates optimal performance when S-1 and S-2 are both employed as the vision encoder, at which point the vision encoder is capable of learning an effective preliminary representation of pedestrians.

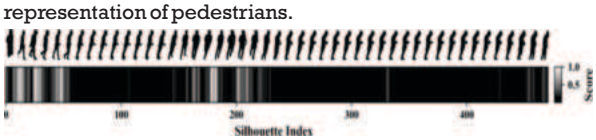

Fig. 5. Silhouette Score in Temporal Max Pooling Phase, Where the Sequence Contains 474 Silhouettes from Gait3D.

Utilizing only S-1 as the vision encoder fails to obtain a satisfactory preliminary pedestrian representation, diminishing the model's learning efficiency within the GL-3D block. Conversely, incorporating S-1, S-2, and S-3 as the vision encoder does not afford additional space for the GL-3D block, impeding the model's ability to learn an effective global temporal receptive field and consequently degrading model performance. Finally, we employ S-1 and S-2 as the vision encoder.

Vision Encoder Component: We also exhibit the component ablation experiments on the vision encoder in Table7. When employing P3D block [16, 36] as the component, GLGait obtains a better result with fewer parameters compared with 3D block [16, 36].

Table VI Ablation Study Of Vision Encoder Components On Gait3dWith Rank-1 Accuracy (%) And Params (M).

Method	Components	Rank-1	Params
GLGait-B	2D block [16]	72.4	3.52
GLGait-B	3D block [16, 36]	73.1	4.12
GLGait-B	P3D block [16, 36]	73.9	3.58
GLGait-L	2D block [16]	75.2	14.07
GLGait-L	3D block [16, 36]	76.4	16.43
GLGait-L	P3D block [16, 36]	77.6	14.28

The possible reason lies in that our GL-3D block also separates the spatial and temporal dimensions, which is similar to P3D block. Maintaining such a similar structure assists in model training. Meanwhile, for 2D block [16], although it has fewer parameters, it is unable to process temporal information, which is essential in pedestrian representation, thus the performance significantly drops out. Finally, we select P3D block as the component in the vision encoder to obtain a good accuracy and cost trade-off.

Effectiveness of PGTA: To verify the effectiveness of Pseudo Global Temporal Self- Attention (PGTA), we compare it with other multi-head self- attention [44] methods, containing Spatio-Temporal MHSA [2], Factorised self-attention [2] on temporal dimension, and MobileViT [33] self-attention. Specifically, we set patch size to 3 × 4.

Table VII Ablation Study Of Memory And Computation Complexity In Self-attention On Gait3d With Rank-1 Accuracy (%), Params (M), And Flops (G).

Method	Module	Rank-1	Params	FLOPs
GLGait-B	Spatio-Temporal MHSA [2]	68.2	7.93	0.93
GLGait-B	Factorised self-attention [2]	70.6	7.93	0.92
GLGait-B	MobileViT self-attention	72.0	3.58	0.94
GLGait-B	PGTA	73.9	3.58	0.87

The results are shown in Table 8. PGTA reduces half of the parameters compared with Spatio-Temporal MHSA and Factorised self-attention. Meanwhile, we observe that the

Rank- 1 accuracy of Spatio-Temporal MHSA and Factorised self- attention greatly drops out. The possible reason is that a large information loss occurs between the 3,072 token size (patch size × channels) and 256 channels. Compared with MobileViT self-attention, PGTA improves 1.9% Rank-1 accuracy with fewer FLOPs. Due to the issue of the receptive field lying in the temporal dimension, PGTA only focuses on the temporal dimension and separates the spatial dimension from tokens, thus effectively establishing a good solution. Effectiveness of Temporal Convolution after PGTA: To verify the effectiveness of temporal convolution after PGTA, we compare it with a normal linear operation. As shown in Table 9, employing temporal convolution improves GLGait-B 1.6% and GLGaitL 1.1% Rank-1 accuracy than a normal linear operation with few parameters increase. Temporal convolution enhances the local receptive field, assisting the model in learning the motion process of gait. Besides, temporal convolution can also aggregate pseudo global temporal receptive fields generated by PGTA to a true holistic temporal receptive field. Its effectiveness is well demonstrated.

Table VIII Ablation Study Of Temporal Convolution After Pgta On Gait3d With Rank-1 Accuracy (%), And Params (M).

Method	Temporal Convolution	Rank-1	Params
GLGait-B	✓	72.3	3.32
GLGait-B	✗	73.9	3.58
GLGait-L	✓	76.5	13.23
GLGait-L	✗	77.6	14.28

E. Visualization

To verify the effectiveness of GLGait in long sequences, we conduct visualization as illustrated in Figure 5, where the score is model attention in temporal max pooling phase for each silhouette. GLGait can detect dynamic sub-sequences and give them high scores; for static sub-sequences, it selects representative silhouettes to give high scores and assigns low scores to the rest. This demonstrates that GLGait can align various gait patterns in long sequences, thus validating the effectiveness of global-local temporal receptive field.

V. CONCLUSION

In this paper, we have introduced a creative way to deal with address the difficulties of transient responsive fields in walk acknowledgment in genuine world, uncontrolled conditions. Perceiving the intricacy and changeability of stride designs in such conditions, we presented a clever organization engineering called the Worldwide Neighborhood Fleeting Responsive Field Organization (GLGait).

This organization influences multi-head self-consideration (MHSA) systems put before the worldly convolution activity in Convolutional Brain Organizations (ConvNets), expecting to catch both worldwide and neighborhood fleeting conditions in step successions. By coordinating MHSA, our model can zero in on huge transient elements and communications across step outlines, which is pivotal for perceiving people on foot under differing conditions. Be that as it may, the utilization of MHSA acquaints difficulties related with high memory and computational expenses because of the dimensionality blast while managing long successions. To alleviate this issue, we propose the Pseudo Worldwide Fleeting Self-Consideration (PGTA) procedure. PGTA effectively lessens the memory and computational intricacy related with MHSA while keeping up with its capacity to catch rich transient conditions. This empowers GLGait to scale really and work with long stride groupings in enormous datasets, making it computationally plausible for organization in commonsense reconnaissance frameworks. Moreover, we presented the Middle Increased Trio Misfortune (CTL), an original misfortune capability intended to upgrade the growing experience of step acknowledgment models. The CTL successfully lessens intra-class distances, guaranteeing that comparative person on foot walk arrangements are all the more firmly assembled in the component space. Simultaneously, it grows the positive example space, which works on the model's capacity to sum up across various walkers. This misfortune capability gives a consistent replacement to conventional trio misfortune, offering better combination and further developed execution in the acknowledgment task. The proposed GLGait organization, with its effective utilization of transient responsive fields, decreased computational intricacy, and upgraded misfortune capability, is especially appropriate for walk acknowledgment in uncontrolled, genuine situations. Its capacity to offset exactness with effectiveness makes it a promising possibility for huge scope observation frameworks, where both execution and asset limits should be thought of.

Broad trials were directed on two testing walk datasets: Gait3D and Developed. These datasets, which contain different and complex stride groupings caught in genuine circumstances, act as benchmarks for assessing walk acknowledgment techniques in nature. The trial results exhibit that our methodology essentially beats cutting edge techniques concerning acknowledgment exactness. In addition to the fact that GLGait achieves unrivaled execution, however it likewise does as such with lower computational and memory necessities, which are basic elements for down to earth sending in observation frameworks. All in all, the GLGait approach presents a promising answer for the difficulties of stride acknowledgment in uncontrolled conditions. By presenting a clever design that successfully catches both worldwide and neighborhood fleeting elements, combined with an effective self-consideration instrument and an upgraded misfortune capability, we have exhibited the capability of GLGait to further develop step acknowledgment exactness while keeping up with computational proficiency. This work opens the entryway for the reception of stride acknowledgment frameworks in certifiable applications, for example, public observation, security checking, and robotized person on foot recognizable proof frameworks.

REFERENCES

[1] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multiview very large population database with pose sequences," IEEE Trans. Biometrics, Behavior, and Identity Science, vol. 2, no. 4, pp. 421–430, 2020.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Luc˘ic´, and C. Schmid"Vivit: A video vision transformer," in Proc. IEEE/CVF Int. Conf., Comput. Vis., 2021, pp. 6836–6846.

[3] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2022, pp. 20249–20258.

[4] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8126–8133.

[5] P. Connor and A. Ross, "Biometric recognition by gait: A survey of modalities and features," Comput. Vis. Image Understand., vol. 167, pp. 1–27, 2018.

[6] M. Deng, Z. Fan, P. Lin, and X. Feng, "Human gait recognition based on frontal-view sequences using gait dynamics and deep learning," IEEE Trans. Multimedia, vol. 26, pp. 117–126, 2024.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkor- eit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Representations, 2021.

[8] H. Dou, P. Zhang, W. Su, Y. Yu, and X. Li, "Metagait: Learning to learn an omni sample adaptive representation for gait recognition," in Proc Eur. Conf. Comput. Vis., 2022, pp. 357–374.

[9] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "GaitGCI: Generative counterfactual intervention for gait recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 5578–5588.

[10] C. Fan, S. Hou, Y. Huang, and S. Yu, "Exploring deep models for practical gait recognition," arXiv preprint arXiv: 2303.03301, 2023.

[11] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "OpenGait: Re- visiting gait recognition towards better practicality," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 9707–9716.

[12] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, "Gaitpart: Temporal part-based model for gait recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14225-14233.

[13] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "GPGait: Generalized pose-based gait recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 19595–19604.

[14] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8295–8302.

[15] M. Geva, A. Caciularu, K. R. Wang, and Y. Goldberg, "Transformer feed-

forward layers build predictions by promoting concepts in the vocabulary space," in Proc. Conf. Empirical Methods Natural Lang. Process., 2022, pp. 30–45.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
[17] X. He, Y. Zhou, Z. Zhou, S. Bai, and X. Bai, "Triplet-center loss for multi-view 3D object retrieval," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1945-1954.
[18] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," arXiv preprint arXiv: 1703.07737, 2017.
[19] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 382–398.
[20] X. Huang, D. Zhu, H. Wang, X. Wang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 12909–12918.
[21] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., "Segment anything," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2023, pp. 4015-4026
[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp.2278-2324, 1998.
[23] A. Li, S. Hou, Q. Cai, Y. Fu, and Y. Huang, "Gait recognition with drones: A benchmark," IEEE Trans. Multimedia, vol. 26, pp. 3530-3540, 2024.
[24] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao, "Transgait: Multimodal- based gait recognition with set transformer," Appl. Intell., vol. 53, no. 2, pp. 1535–1547, 2023.
[25] N. Li and X. Zhao, "A strong and robust skeleton-based gait recognition method with gait periodicity priors," IEEE Trans. Multimedia, vol. 25, pp 3046-3058, 2023.