Seeing Wider, Seeing Longer: OmniGait for Gait Recognition in the Wild with Re-parameterized Spatio-Temporal Large Kernel Convolution

Kewen Dong[†]
College of Software
Hebei Normal University
Shijiazhuang, China
kewen_dong@163.com

Zuhan Fan[*]
College of Software
Hebei Normal University
Shijiazhuang, China
zuhan_fan@163.com

Yitong Li[*]
College of Software
Hebei Normal University
Shijiazhuang, China
yitong_li2005@163.com

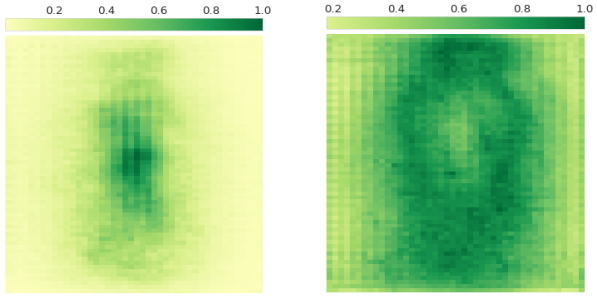
Pan Dou[†]
College of Software
Hebei Normal University
Shijiazhuang, China
doupan18@163.com

***Abstract*—Gait recognition, as a long-range, non-invasive human identification technology, has garnered increasing attention because of its convenience and effectiveness. Compared to methods that achieve significant results in laboratory settings, gait recognition methods in outdoor scenarios face a more urgent need for robust morphological feature extraction and effective modeling of long-range gait sequences. To capture more discriminative gait features from outdoor data, this paper proposes OmniGait, a method based on reparameterized large kernel convolution. Specifically, this method innovatively proposes a spatio-temporal large kernel convolutional module(ST-LK), which incorporate two decoupled branches: spatial and temporal. The spatial branch leverages the large effective receptive field(ERF) of 2D large kernel convolution to capture more discriminative pedestrian morphological features from sparse contour data. The temporal branch employs efficient 1D convolution to model long gait sequences, enabling the model to acquire near-global temporal receptive fields. This allows it to directly capture long-term motion patterns spanning dozens of frames. Additionally, this paper proposes a dynamic gating mechanism(DGM) that enables the spatio-temporal large kernel convolutional module(ST-LK) to adaptively adjust the weights of convolutional branches of different sizes based on the input data. This approach enhances the robustness of the gait features extracted by the model. Extensive experiments demonstrate that our proposed method achieves outstanding results on both the popular outdoor gait datasets Gait3D and GREW. The source code will be available.**

***Index Terms*—Gait Recognition, Effective Receptive Field, Spatio-Temporal Large Kernel Convolutional Module, Dynamic Gating Mechanism**

I. Introduction

Gait recognition, as an emerging biometric technology, aims to authenticate individuals by extracting unique biomechanical characteristics embedded in their walking patterns through deep learning networks [1]–[3]. Currently, many appearance-based gait recognition methods [4]–[7] have achieved remarkable results on controlled laboratory environment datasets [8], [9]. However, when transferring to outdoor real-world datasets [10], [11], the performance of existing methods often experiences a steep decline. The root cause of this performance gap lies in the complex environmental noise present in the real world, which makes it difficult for models to decouple effective identity features from disturbed data. To address the impact of real-world environmental noise on gait data, this paper examines the issue from both spatial and temporal perspectives.



(a) DeepGaitv2 (b) OmniGait

Fig. 1. Visualization of the model's ERF

For spatial dimensions, large kernel convolution have been shown to exhibit stronger shape bias [12], [13]. As shown in Figure 1, compared to small-sized convolutional neural networks, large spatial kernels can more effectively cover the information-dense pedestrian contour regions within gait contour maps through their extensive spatial receptive fields. This enables the capture of more discriminative morphological structures of pedestrians, thereby mitigating noise interference from occlusions [14], background variations, and arbitrary viewing angles. For temporal dimensions, the random walking speeds of pedestrians in real-world scenarios and fragmented gait sequences are constrained by the limited temporal receptive fields of traditional small-sized 3D CNNs, making it difficult to capture complete gait cycles. The key to solving this problem lies in introducing the concept of large kernel convolution in the temporal dimension. This approach enables the direct establishment of a global temporal receptive field that spans nearly the entire gait cycle at minimal computational cost, thereby effectively modeling movement patterns in long

[†] Corresponding author.
[*] Contributed equally to this work as co-first authors.

sequences.

Based on the insights into spatio-temporal dimensions, this paper proposes OmniGait, an efficient gait recognition network based on reparameterized spatio-temporal large kernel convolutional module. The core of the method lies in the innovative construction of a spatio-temporal large kernel convolutional module(ST-LK). This module draws inspiration from the decoupling concept of pseudo-3D convolution [1], [2]. We orthogonally decouple computationally expensive 3D large kernel convolution across spatial and temporal dimensions: the spatial large kernel branch focuses on capturing robust morphological features of pedestrians from each frame using an enormous spatial receptive field; while the temporal kernel branch concentrates on constructing an ultra-long temporal effective receptive field across the sequence dimension to capture the pedestrian's movement rhythms and periodic patterns from a global perspective. Finally, through an efficient feature fusion module, complementary features of appearance and motion are organically integrated to generate a highly discriminative representation of pedestrian identity.

Additionally, to further enhance the model's robustness in uncontrolled random environments, we designed a dynamic gating mechanism(DGM) for the ST-LK. In outdoor settings, pedestrian silhouettes undergo significant changes, and the length of effective gait sequences exhibits high uncertainty. Convolutional kernels with fixed weights struggle to adapt to these diverse distributions simultaneously. Therefore, we believe that mechanisms enabling models to adaptively adjust based on input data are essential. This DGM can perceive the scale and temporal characteristics of input data, adaptively adjusting the activation weights of convolutional branches with different sizes within the large kernel convolutional module. This enables OmniGait to dynamically focus on the most effective feature scale when encountering pedestrians of diverse shapes or gait sequences of varying lengths, thereby achieving robust feature extraction. The main contributions of this paper are as follows:

- This paper innovatively proposes a spatio-temporal large kernel convolutional module. It employs a spatio-temporal decoupling strategy to simultaneously extract robust pedestrian morphological features and capture complete pedestrian motion patterns, thereby obtaining richer representations of pedestrian identity.
- This paper proposes a novel dynamic gating mechanism that enables the model to adaptively adjust the weights of convolutional branches based on data, thereby extracting more robust features.
- This paper constructs the efficient gait recognition network OmniGait based on a reparameterized spatio-temporal large kernel convolutional module, enhancing the model's feature extraction capabilities in outdoor environments from a spatio-temporal perspective.
- We conducted extensive experiments on large-scale outdoor datasets Gait3D and GREW, thoroughly validating OmniGait's effectiveness and superiority in real-world outdoor environments.

II. Related Work

A. Appearance-Based Gait Recognition

Appearance-based gait recognition primarily falls into two categories: set-based methods and sequence-based methods. Set-based methods such as GaitSet [4] and GaitSSB [15] treat gait as an unordered set for feature aggregation. While robust to viewpoint changes, they overlook critical temporal patterns. Sequence-based methods explicitly model spatio-temporal dependencies: GaitPart [5] focuses on local micro-motion features, while GaitGL [6] and MTSGait [10] enhance spatio-temporal information capture through improved convolutional modules. DyGait [16] further decouples dynamic and static features. Recently, DeepGaitV2 [17] has validated the effectiveness of deep neural networks in outdoor scenarios. However, the aforementioned methods largely rely on stacking small-sized convolutional kernels (such as 3×3). When processing sparse and low-quality outdoor binary contours, they are constrained by their small effective receptive fields, making it difficult to extract robust morphological features. Moreover, capturing long-range temporal dependencies is equally challenging in outdoor settings where gait sequences vary in length and lack alignment. Traditional 3D CNNs are constrained by their local receptive fields, making it difficult to cover the entire gait cycle. Unlike the aforementioned methods, the OmniGait approach proposed in this paper achieves modeling of spatial morphology and temporal long-range dependencies through decoupled ST-LK, while maintaining computational efficiency.

B. Large Kernel Convolutional Neural Networks

To bridge the gap between Convolutional Neural Networks and Vision Transformers [18] in visual tasks, ConNeXt [19] introduces 7×7 convolutional kernels to grant the model a larger effective receptive field, thereby achieving significant performance improvements. Subsequently, RepLKNet [20] employs structural reparameterization to expand the convolutional kernel size to 31×31, yielding a denser and broader effective receptive field that enables the model to better capture shape features across spatial dimensions. In subsequent research, SLak [21] combined two rectangular convolutional kernels into a larger kernel based on convolutional decomposition, while UniRepLKNet [22] enhanced its ability to capture sparse spatial patterns through dilation reparameterization. Additionally, PeLK [23] proposes a parameter-efficient peripheral convolution mechanism that achieves efficient computation for an ultra-large 101×101 kernel through dynamic focusing and fuzzy region segmentation. However, how to effectively extend the advantages of large kernel convolution from the spatial dimension to the temporal dimension, leveraging their long-range modeling capabilities to address long-sequence dependencies in uncontrolled scenarios, remains an area demanding urgent exploration.

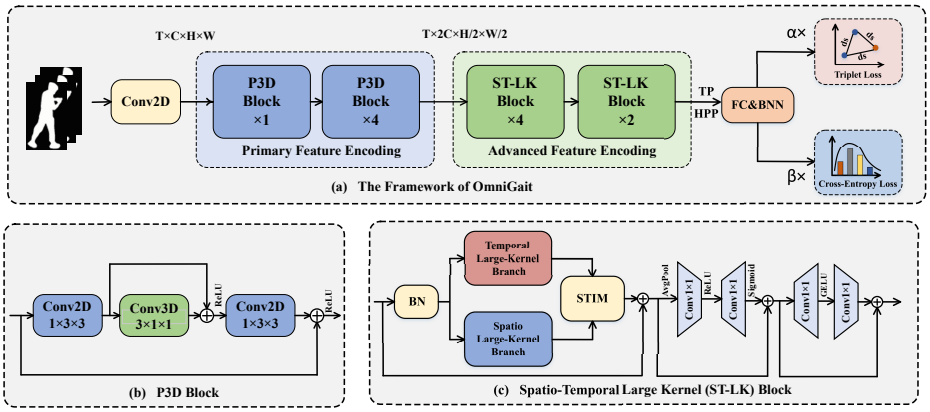


Fig. 2. The framework of OmniGait

III. Method

A. Overview

As shown in Figure 2, the overall architecture of OmniGait follows a hierarchical feature extraction approach and a design philosophy that progresses from coarse to fine. The input contour sequence first undergoes feature densification processing on the sparse binary silhouette map through an initial 2D convolutional layer. The resulting feature matrix then undergoes a primary feature encoding stage to obtain a primary feature representation. During the high-level feature encoding stage, ST-LK are employed to obtain advanced spatio-temporal representations that balance both global and local information. Subsequently, feature aggregation is performed using temporal pooling and horizontal pyramid pooling. Finally, we conduct end-to-end optimization training using a joint loss function composed of triplet loss and cross-entropy loss, with the specific formula expressed as follows:

$$L_{total} = \alpha L_{tri} + \beta L_{ce} \tag{1}$$

Where, α and β serve as hyperparameters representing the respective weights of the two loss functions, thereby balancing the overall loss.

B. Primary Feature Encoder

This paper proposes a primary feature encoder as a preprocessing module that preliminarily extracts local spatio-temporal cues while downsampling feature maps to an appropriate scale. This enables subsequent high-level encoders to focus more effectively on modeling global spatio-temporal dependencies. The design employing a P3D convolutional module to construct the primary encoder significantly reduces the model's parameter count and computational complexity while preserving its ability to model spatio-temporal context. Specifically, the input $X \in R^{N\times C\times T\times H\times W}$ sequentially passes through two 2D convolution and one 1D convolution at this stage. The input features and the processed spatio-temporal features are then added together via residual connections.The entire process is specifically represented as follows:

$$x_1 = ReLU(Conv_{2d}(x_{in})) \tag{2}$$

$$x_2 = Conv_{1d}(x_1) \tag{3}$$

$$x_3 = Conv_{2d}(ReLU(x_1 + x_2)) \tag{4}$$

$$x_{out} = ReLU(x_{in} + x_3)) \tag{5}$$

C. Spatio-Temporal Large Kernel Convolutional Module

To enhance the spatio-temporal feature extraction capability of gait recognition models in outdoor scenarios while maintaining a balance between extraction efficiency and cubic computational complexity, we propose a ST-LK as shown in Figure 3 based on an orthogonal decoupling strategy, drawing inspiration from P3D convolution. We decouple the expensive $K \times K \times K$ spatio-temporal operation into parallel $1 \times K \times K$ spatial large kernel branches and $K \times 1 \times 1$ temporal large kernel branches. This design forces the model to focus on learning spatial patterns and modeling temporal rhythms in separate branches. Not only does it eliminate mutual interference between spatio-temporal features, but it also achieves feature extraction capabilities comparable to or even superior than standard 3D large kernel convolution while maintaining low computational complexity.

The ST-LK primarily consists of parallel spatio-temporal large kernel branches, feature fusion units, and spatio-temporal attention modules. First, to enhance interactions between feature channels while controlling computational cost, we map feature maps to an intermediate dimension through batch normalization and pointwise convolution. This process can be expressed as:

$$X_{proj} = ReLU(BN(Conv(BN(X_{in})))) \tag{6}$$

Among these, X_{proj}represents the preprocessed features after channel compression and nonlinear transformation, which will be fed into subsequent parallel branches.The parallel spatio-temporal core branch, as the core component of the ST-LK, employs a dual-channel parallel architecture to orthogonally decouple spatio-temporal features.

The spatial large kernel branch focuses on extracting robust morphological features within a single frame image. We utilize 2D large kernel convolution that slide across the spatial dimension while remaining independent in the temporal dimension.

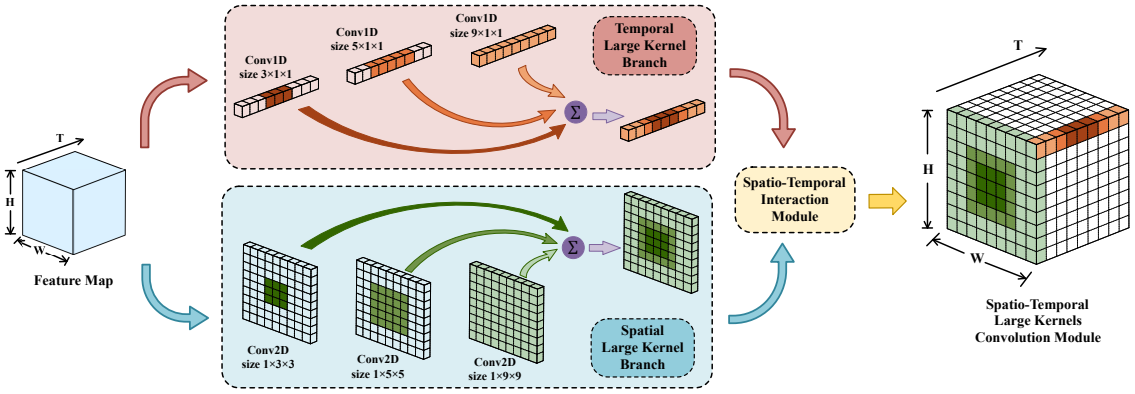


Fig. 3. The method of spatio-temporal large kernel convolutional module(ST-LK)

The computational process can be formalized as a dynamic weighted sum of outputs from multiple spatial convolution kernels:

$$\mathbf{F}_s = \sum_{i=1}^{M} \mathcal{G}_s^{(i)}(\mathbf{X}_{pre}) \odot \mathrm{Conv}_{2d}^{(i)}(\mathbf{X}_{pre}) \tag{7}$$

Where, F_s denotes the spatial features of the output; M represents the number of convolutional kernels within a branch;$Conv_{2d}^{(i)}$) denotes the operation of the i-th convolutional kernel, whose kernel size is strictly constrained to $1 \times K_i \times K_i$ to ensure that the receptive field is established solely in the spatial dimension. $\mathcal{G}_s^{(i)}(\cdot)$ is a dynamic gating mechanism (see Section 3.4 for details) used to generate spatial weight maps corresponding to the input content; $\odot$ denotes the Hadamard product. Through this mechanism, the model can adaptively focus on either the complete pedestrian silhouette or localized edge details.

The temporal large kernel branch is dedicated to establishing long-range dependencies across the temporal dimension. Orthogonal to the spatial branch, we employ one-dimensional large kernel convolution that slide along the temporal dimension while maintaining a 1×1 kernel in the spatial dimension. This enables the model to directly capture motion patterns spanning dozens of frames. Its mathematical expression is:

$$\mathbf{F}_t = \sum_{j=1}^{M} \mathcal{G}_t^{(j)}(\mathbf{X}_{pre}) \odot \mathrm{Conv}_{1d}^{(j)}(\mathbf{X}_{pre}) \tag{8}$$

Where, F_t denotes the temporal feature of the output; $\mathrm{Conv}_{1d}^{(j)}$ represents a temporal convolution operation with kernel size $K_j \times 1 \times 1$. This design endows the model with a near-global temporal receptive field, enabling it to effectively recover complete gait cycle information from uncontrolled, fragmented sequences while avoiding interference with spatial features.

To reintegrate the decoupled features, we designed an efficient spatio-temporal interaction module. Unlike simple element-wise addition, we first concatenate spatial and temporal features along the channel dimension. Subsequently, we employ 1×1 convolution to aggregate and compress cross-channel information, enabling the model to adaptively learn optimal combination weights for shape and motion information:

$$\mathbf{F}_{fused} = \mathrm{Conv}_{mix}([\mathbf{F}_s, \mathbf{F}_t]) \tag{9}$$

Finally, to further enhance feature discriminative power, we feed the fused features into a 3D Squeeze-and-Excitation module. This module learns inter-channel dependencies by compressing global spatio-temporal information, explicitly recalibrates channel weights, suppresses noise responses, and enhances the expression of key gait features. The final output is then added to the input features via residual connections:

$$\mathbf{X}_{out} = \mathbf{X}_{in} + \mathrm{Conv}_{out}(SE(\mathbf{F}_{fused})) \tag{10}$$

D. Dynamic Gating Mechanism

In real-world outdoor surveillance scenarios, to endow the model with the ability to perceive input content and adaptively adjust its receptive field, we embedded a lightweight DGM within the ST-LK. As shown in Figure 4, this mechanism aims to dynamically generate channel-level attention weights for each convolutional branch based on the global contextual information of input features, thereby achieving feature aggregation tailored to individual characteristics.

Specifically, assume the input feature of the dynamic gate controller is $X \in R^{N \times C \times \cdots}$ (a 4D tensor in the spatial branch and a 5D tensor in the temporal branch). The process of generating gating weights comprises the following steps: global context modeling, gating weight generation, and dynamic feature aggregation. First, to capture the global distribution

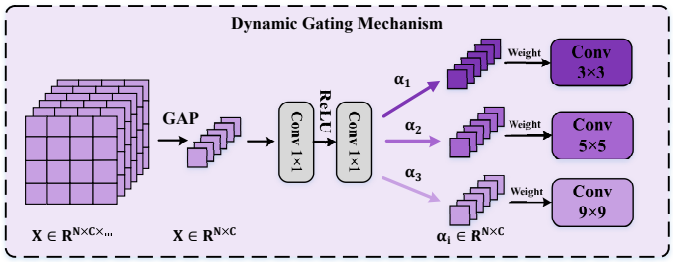


Fig. 4. The framework of Dynamic Gating Mechanism(DGM)

information of the input at minimal computational cost, we employ Global Average Pooling (GAP) to compress features from the spatial or temporal dimension into a global context vector $\mathbf{Z} \in \mathbb{R}^{N \times C}$, thereby aggregating the global response intensity of that channel. This aggregates the global response intensity of the channel. For spatial input X, its calculation formula is expressed as:

$$\mathbf{z}_c = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbf{X}_{c,h,w} \tag{11}$$

Subsequently, to capture nonlinear dependencies between channels and generate multi-branch weights, we feed the context vector z into a unit incorporating a bottleneck structure. This unit consists of two 1×1 convolutional layers separated by a ReLU activation function. The first convolutional layer reduces the channel dimension from C to C/r, while the second convolutional layer expands the dimension to M×C, where M represents the number of parallel convolutional branches. Finally, the output is normalized to the interval (0, 1) via the sigmoid function to generate the final gated weight matrix W, expressed as follows:

$$W = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \tag{12}$$

Here, δ denotes the ReLU function, σ denotes the Sigmoid function, $\mathbf{W}_1 \in R^{\frac{C}{r} \times C}$and $\mathbf{W}_2 \in R^{(M \cdot C) \times \frac{C}{r}}$represent the weight parameters for the two convolutional layers, respectively. The generated W is subsequently reshaped into an $N \times M \times C \times 1 \times \ldots$ matrix, representing the importance coefficient $\alpha_{i,c}$ of the i-th branch on the c-th channel.

The convolutional outputs $\widetilde{\mathbf{F}}_i$ from each branch undergo a channel-wise weighted sum based on the generated gated weights, yielding the adaptively aggregated output feature $\mathbf{F}_{out}$.

$$\mathbf{F}_{out} = \sum_{i=1}^{M} \alpha_i \odot \widetilde{\mathbf{F}}_i \tag{13}$$

Here, $\alpha_i \in R^{N \times C}$ corresponds to the weight vector of the i-th branch, and $\odot$ denotes element-wise multiplication under the broadcast mechanism.

This design enables the ST-LK to dynamically adjust the contribution ratio between large and small kernels based on the real-time characteristics of the input samples. This dynamic aggregation strategy significantly enhances the model's feature robustness and generalization capability in uncontrolled outdoor environments.

IV. Experiments

A. Datasets

The Gait3D dataset is a highly representative real-world gait recognition dataset featuring multi-camera, multi-view, multi-trajectory, and heavily occluded scenes. It comprises 4,000 subjects and 25,309 gait sequences. This dataset divides the top 3,000 pedestrians into a training set and the remaining 1,000 into a test set based on identity, comprehensively covering factors such as common lighting variations, differences in walking directions, and fluctuations in human scale found in outdoor environments.

GREW is currently the largest and most complex outdoor gait recognition dataset, collected across multiple regions using 882 real surveillance cameras. It comprises 26,345 pedestrians and 128,671 sequences. The dataset is divided into training, validation, and test sets. The training set contains sequences from 20,000 pedestrians, the validation set contains sequences from 345 pedestrians, and the test set contains 24,000 sequences from 6,000 pedestrians. The dataset encompasses multiple challenging factors including cross-camera spatiotemporal shifts, random occlusions, long-range views, and low-light conditions.

B. Implementation Details

This implementation uses gait contour sequences as model input during both training and testing phases. Contours undergo normalization preprocessing to adjust them to a 64x44 pixel size. For the datasets Gait3D and GREW used in the experiment, the batch size was set to [32,4], indicating 32 objects, each with 4 sequences. During the training phase, we employed a dynamic sampling approach for frame selection. Specifically, for the Gait3D dataset, we randomly sampled 10–50 frames per sequence, while for the GREW dataset, we randomly sampled 20–40 frames per sequence. During the testing phase, all available frames are input into the model for identity inference. To prevent excessive memory overhead, each sequence is limited to 720 frames.

We employed the stochastic gradient descent (SGD) optimizer to optimize the model. We adjusted different decay strategies to accommodate various datasets. On the Gait3D dataset, we trained the model for 80k iterations with an initial learning rate of 0.1, decaying the learning rate by 0.1 at (30K, 50K, 65K). On the GREW dataset, the model underwent 180K iterations with an initial learning rate of 0.05. The learning rate was decayed at (60K, 120K, 150K) with a decay rate of 0.2. To accelerate model convergence and prevent getting stuck in local optima, momentum is set to 0.9; the weight decay coefficient is set to 5e-4 to avoid weight explosion. This paper applies random spatial augmentation to the data: random perspective probability. The random horizontal flipping probability and random rotation probability are both set to 0.2.

C. Comparison with State-of-the-Art Methods

We conducted a comprehensive performance evaluation of the proposed method on the Gait3D dataset. Table I presents the detailed experimental results. Overall, our approach consistently demonstrates superior performance among all state-of-the-art methods evaluated. Notably, our model achieves a 12.4% higher Rank-1 accuracy than the current benchmark model, OpenGait. Furthermore, compared to DeepGait, which previously achieved the best results, our method outperforms it by 2.6% and 4.6% in Rank-1 and mAP accuracy, respectively. These compelling results significantly highlight the effectiveness of our proposed approach.

TABLE I
COMPARISON WITH NUMEROUS STATE-OF-THE-ART METHODS ON GAIT3D AND GREW

Method	Publication	Gait3D				GREW			
		Rank-1	Rank-5	mAP	mINP	Rank-1	Rank-5	Rank-10	Rank-20
GaitPart [5]	CVPR 2020	28.2	47.6	21.6	12.4	44.0	60.7	67.3	73.5
GaitGL [6]	ICCV 2021	29.7	48.5	22.3	13.3	47.3	63.6	69.3	74.2
GaitSet [4]	TPAMI 2021	36.7	58.3	30.0	17.3	46.3	63.6	70.3	76.8
MTSGait [10]	ACMMM 2022	48.7	67.1	37.6	21.9	55.3	71.3	76.9	81.6
TAG-GS [24]	TIM 2025	54.9	73.7	45.9	28.6	-	-	-	-
GaitGCI [25]	CVPR 2023	57.2	74.5	45.0	27.6	68.5	80.8	84.9	-
HSTL [26]	ICCV 2023	61.3	76.3	55.5	-	62.7	76.6	81.3	-
TriGait [27]	TBIOM 2024	61.4	79.4	52.4	29.8	-	-	-	-
GaitSSB [15]	TPAMI 2023	63.6	-	-	-	61.7	-	-	-
GaitBase [28]	CVPR 2023	64.6	79.8	53.4	31.4	60.1	75.5	80.4	84.1
DyGait [16]	ICCV 2023	66.3	80.8	56.4	37.3	71.4	83.2	86.8	89.5
QAGait [29]	AAAI 2024	67.0	81.5	56.5	-	59.1	74.0	79.2	83.1
CLTD [30]	ECCV 2024	69.7	85.2	67.2	-	78.0	87.8	-	-
DeepGaitV2 [17]	ArXiv 2023	74.4	88.0	65.8	-	77.7	87.9	90.6	-
OmniGait	Ours	**77.0**	**89.5**	**70.4**	**52.3**	**81.5**	**89.7**	**92.3**	**94.2**

TABLE II
ABLATION STUDY ON KERNEL SIZES. S/T DENOTE SPATIAL/TEMPORAL.

S-Large Kernel	S-Small Kernel	T-Large Kernel	T-Small Kernel	Rank-1	mAP
3×3	-	$3 \times 1 \times 1$	-	74.4	65.8
5×5	3×3	$5 \times 1 \times 1$	$3 \times 1 \times 1$	75.5	66.8
7×7	3×3 5×5	$7 \times 1 \times 1$	$3 \times 1 \times 1$ $5 \times 1 \times 1$	76.1	68.9
9×9	3×3 5×5	$9 \times 1 \times 1$	$3 \times 1 \times 1$ $5 \times 1 \times 1$	**77.0**	**70.4**
13×13	3×3 5×5	$13 \times 1 \times 1$	$3 \times 1 \times 1$ $5 \times 1 \times 1$	76.6	69.5

TABLE III
ABLATION STUDY OF DIFFERENT COMPONENTS IN THE PROPOSED FRAMEWORK.

Model Variant	**S-LK**	**T-LK**	**DGM**	**Rank-1**	**mAP**
Baseline (Small Kernel)	×	×	×	74.4	65.8
Spatial Large Kernel Only	✓	×	×	75.8	67.5
Temporal Large Kernel Only	×	✓	×	76.3	69.2
OmniGait (Ours)	✓	✓	✓	**77.0**	**70.4**

Performance on the GREW dataset further demonstrates the effectiveness of our method. Compared to all methods listed in Table III, OmniGait exhibits more convincing performance. Specifically, OmniGait achieves a 21.4% higher Rank-1 accuracy than the baseline method OpenGait. Furthermore, it surpasses the current state-of-the-art method DeepGait by 3.8% in Rank-1 accuracy. These results demonstrate that modeling more diverse and comprehensive spatiotemporal receptive fields is highly effective for achieving robust gait recognition when challenged with real-world datasets, further highlighting our model's superior fitting performance on outdoor gait data sourced from the real world.

D. Ablation Experiment

To validate the effectiveness of each core design in OmniGait, we conducted extensive ablation experiments on the Gait3D dataset, with all experiments employing the same training configuration.

1) Optimal Structure for Large Kernel Convolution: To investigate the impact of large kernel convolution sizes on model performance across spatial and temporal dimensions. Larger kernel convolution provide a broader receptive field, but excessively large kernels may cause overfitting or feature smoothing on low-resolution gait images (64×44). As shown in Table II, increasing the size of the large kernel convolution from 3×3 to 9×9 significantly improves the model's Rank-1 accuracy. This demonstrates that larger spatio-temporal effective receptive fields are beneficial for capturing complete pedestrian shapes and long-term motion patterns. However, when the kernel size is further increased to 13×13, performance shows slight saturation or even decline. This may be attributed to the lower resolution of gait contour maps, where excessively large kernel convolution tend to cause over-smoothing of features. Therefore, we ultimately selected 9×9 as the optimal configuration.

2) Effectiveness of Spatio-Temporal Branch Parallel Structures: To investigate whether decoupled the ST-LK demonstrate superiority over models that focus solely on either temporal or spatial dimensions. We designed variant networks that retain only the spatial main branch and only the temporal main branch, respectively, and conducted comparative experiments against the current decoupled parallel spatio-temporal main branch convolutional network. As shown in Table III, feature extraction based solely on global receptive field modeling in a single spatial or temporal dimension cannot achieve optimal performance. Our proposed parallel decoupling architecture

(OmniGait) achieves complementary integration of form and motion by simultaneously leveraging two branches, delivering significant performance improvements over single-branch models. We also introduced variants incorporating dynamic gate mechanisms, demonstrating that the model adaptively adjusts the weights of large and small kernels based on the characteristics of input samples, enabling the extraction of more robust features.

V. Conclusion

This paper proposes an outdoor gait recognition method based on the concept of reparameterized large kernel convolution. This method decouples spatio-temporal large kernel convolutional module. The decoupled spatial kernel branch expands the effective receptive field in the spatial dimension to extract more robust morphological features. Meanwhile, the temporal kernel branch establishes longer temporal dependencies to capture complete motion patterns. The proposed dynamic gating mechanism enables the spatio-temporal large kernel convolutional module to adaptively adjust the weights of convolutional branches of different sizes based on the input data, thereby obtaining more discriminative identity features. At this stage, the impact of multimodal data on gait recognition tasks has not been further explored. Moving forward, we will leverage multimodal gait data to develop more advanced outdoor gait recognition methods.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[2] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5533–5541.

[3] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[4] H. Chao, Y. He, J. Zhang, and J. Feng, "Gaitset: Regarding gait as a set for cross-view gait recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8126–8133.

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 225–14 233.

[6] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 648–14 656.

[7] X. Huang, D. Zhu, H. Wang, X. Wang, B. Yang, B. He, W. Liu, and B. Feng, "Context-sensitive temporal feature learning for gait recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 909–12 918.

[8] S. Yu, D. Tan, and T. Tan, "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)*, vol. 4. IEEE, 2006, pp. 441–444.

[9] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ transactions on Computer Vision and Applications*, vol. 10, no. 1, p. 4, 2018.

[10] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20 228–20 237.

[11] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, and J. Zhou, "Gait recognition in the wild: A benchmark," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 789–14 799.

[12] Y. Zhang and M. A. Mazurowski, "Convolutional neural networks rarely learn shape for semantic segmentation," *Pattern Recognition*, vol. 146, p. 110018, 2024.

[13] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," *Advances in neural information processing systems*, vol. 29, 2016.

[14] Y. Peng, C. Cao, and Z. He, "Occluded gait recognition," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[15] C. Fan, S. Hou, J. Wang, Y. Huang, and S. Yu, "Learning gait representation from massive unlabelled walking videos: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 920–14 937, 2023.

[16] M. Wang, X. Guo, B. Lin, T. Yang, Z. Zhu, L. Li, S. Zhang, and X. Yu, "Dygait: Exploiting dynamic representations for high-performance gait recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 13 424–13 433.

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

[18] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[20] X. Ding, X. Zhang, J. Han, and G. Ding, "Scaling up your kernels to 31x31: Revisiting large kernel design in cnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 963–11 975.

[21] S. Liu, T. Chen, X. Chen, X. Chen, Q. Xiao, B. Wu, T. Kärkkäinen, M. Pechenizkiy, D. Mocanu, and Z. Wang, "More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity," *arXiv preprint arXiv:2207.03620*, 2022.

[22] X. Ding, Y. Zhang, Y. Ge, S. Zhao, L. Song, X. Yue, and Y. Shan, "Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5513–5524.

[23] H. Chen, X. Chu, Y. Ren, X. Zhao, and K. Huang, "Pelk: Parameter-efficient large kernel convnets with peripheral convolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5557–5567.

[24] M. S. Shakeel, K. Liu, X. Liao, and W. Kang, "Tag: A temporal attentivegait network for cross-view gait recognition," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[25] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "Gaitgci: Generative counterfactual intervention for gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5578–5588.

[26] L. Wang, B. Liu, F. Liang, and B. Wang, "Hierarchical spatio-temporal representation learning for gait recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 19 582–19 592.

[27] Y. Sun, X. Feng, X. Liu, L. Ma, L. Hu, and M. S. Nixon, "Trigait: hybrid fusion strategy for multimodal alignment and integration in gait recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 7, no. 1, pp. 82–94, 2024.

[28] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "Opengait: Revisiting gait recognition towards better practicality," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 9707–9716.

[29] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, Y. Huang, P. Li, and S. Xu, "Qagait: Revisit gait recognition from a quality perspective," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5785–5793.

[30] H. Xiong, B. Feng, X. Wang, and W. Liu, "Causality-inspired discriminative feature learning in triple domains for gait recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 251–270.