GaitMA: Pose-guided Multi-modal Feature Fusion for Gait Recognition

Fanxu Min, Shaoxiang Guo, Hao Fan✉, Junyu Dong✉
Faculty of Information Science and Engineering
Ocean University of China
Qingdao, China
{minfanxu, guoshaoxiang}@stu.ouc.edu.cn
{dongjunyu, fanhao}@ouc.edu.cn

***Abstract*—Gait recognition is a biometric technology that recognizes the identity of humans through their walking patterns. Existing appearance-based methods utilize CNN or Transformer to extract spatial and temporal features from silhouettes, while model-based methods employ GCN to focus on the special topological structure of skeleton points. However, the quality of silhouettes is limited by complex occlusions, and skeletons lack dense semantic features of the human body. To tackle these problems, we propose a novel gait recognition framework, dubbed Gait Multi-model Aggregation Network (GaitMA), which effectively combines two modalities to obtain a more robust and comprehensive gait representation for recognition. First, skeletons are represented by joint/limb-based heatmaps, and features from silhouettes and skeletons are respectively extracted using two CNN-based feature extractors. Second, a co-attention alignment module is proposed to align the features by element-wise attention. Finally, we propose a mutual learning module, which achieves feature fusion through cross-attention, Wasserstein loss is further introduced to ensure the effective fusion of two modalities. Extensive experimental results demonstrate the superiority of our model on Gait3D, OU-MVLP, and CASIA-B.**

***Index Terms*—Gait recognition, multi-model, feature fusion, deep neural network**

I. Introduction

Gait recognition has recently gained widespread interest as a biometric technology that recognizes people by their walking patterns. Unlike other biometrics like face, fingerprint, and iris, gait can be captured from a distance in uncontrolled settings without the cooperation of individuals. However, this challenging technique still faces many difficulties, including complex backgrounds, severe occlusion, unpredictable illumination, arbitrary viewpoints, and diverse clothing changes. The appearance-based methods mainly extract temporal and spatial features from silhouettes by 2D/3D CNN, Transformer, RNN, and LSTM [1], [2], [3]. They focus on extracting features from the whole gait sequence or adjacent frames, this makes them perform poorly when facing lower-quality silhouettes. The model-based methods [4], [5], [6], [7], [8] mostly take clear and robust skeletons as the input, skeletons in a video are mainly represented as a sequence of joint coordinates which are extracted by pose estimators [9]. Benefiting from the rapid development in pose estimation and the application of Graph Convolutional Network (GCN) [10], recent model-based methods could even show competitive results compared to appearance-based methods.

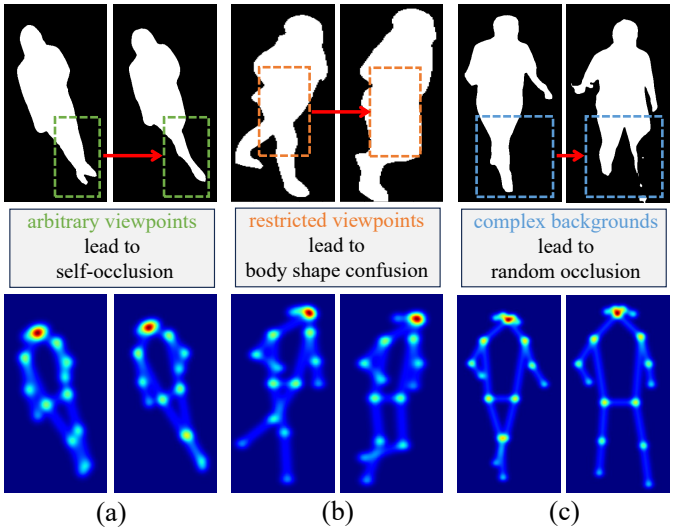

Fig. 1. A brief visualization of our motivation. Skeleton can effectively complement missing gait features in silhouette across various challenging scenarios.

However, modality aggregation in gait recognition is rarely discussed [11]. First, as shown in Fig. 1(a), due to arbitrary viewpoints, the left leg is missing in motion due to self-occlusion, silhouettes can not provide complete gait information in this case, but skeletons give a clear representation of current motion state. Second, due to the problem of self-occlusion in motion, the shape of the human body changes considerably, and it is difficult to distinguish between the torso and the limbs, as shown in Fig. 1(b), skeletons can guide the posture of the human body to obtain a more robust gait representation. Finally, silhouettes are easily obscured by complex backgrounds and lose shape information, as shown in Fig. 1(c), skeletons can complement missing gait features in silhouettes. It can be observed that the silhouette retains the external body shape information and omits some body-structure clues, and the skeleton preserves the internal body structure information. The two data modalities are complementary to each other, but

✉ Corresponding author.

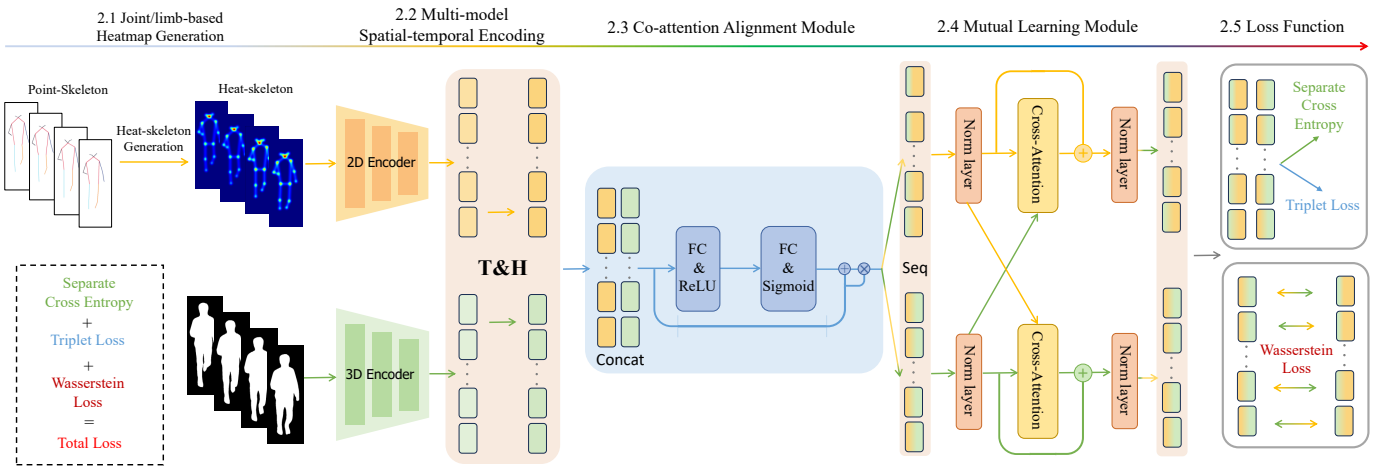

Fig. 2. An overview of the proposed framework GaitMA for gait recognition. T&H represents the horizontal mapping and temporal aggregation. Concat and Seq denote the features concatenate and separate, respectively.

they may not correspond, containing mismatched redundancy and interference information. Therefore, how to better fuse the silhouette and skeleton is a challenging problem, which significantly influences the performance of obtaining a comprehensive representation of gait.

To achieve this goal, we propose a novel gait recognition modality fusion framework, named GaitMA, which effectively combines two modalities to obtain a more robust and comprehensive gait representation for recognition. First, we obtain joint/limb-based heatmaps by computing the Gaussian distribution of skeletal points to enhance the robustness and interoperability of the skeleton [12], [13]. This reduces the modality differences between the skeleton and the silhouette. Subsequently, we built a novel asymmetric CNN-based dual-branch architecture to individually extract spatial-temporal gait features from each modality. Second, to effectively integrate the two modalities and fully utilize their information, the proposed co-attention alignment module is introduced to mitigate feature redundancy and interference. It achieves alignment by calculating feature attention between elements, thereby bringing the feature distributions of the two modalities closer in the feature space. Finally, the mutual learning module is proposed to facilitate the interaction between the two modalities. This module effectively enriches discrete skeleton representations and complements the semantic information of silhouette images. Additionally, Wasserstein loss [14] is introduced to ensure comprehensive mutual learning of features between the two modalities.

The main contributions of the proposed method are summarized as follows: (1) We propose a novel gait recognition modality fusion framework called GaitMA, which utilizes a more comprehensive gait representation constructed from both silhouettes and skeletons represented by joint/limb-based heatmaps to achieve better recognition performance. (2) A co-attention alignment module is proposed to improve the efficiency and effectiveness of feature interaction. (3) We propose a mutual learning module for feature fusion and introduce Wasserstein loss to ensure effective fusion of the two modalities. Experimental results demonstrate that our method achieves superior performance on three dominant datasets, it obtains an average Rank-1 accuracy of 66.1% on Gait3D, 95.9% on CASIAB, and 91.2% on OU-MVLP, respectively.

II. METHOD

In this section, we will describe the specific details of the model implementation. As shown in Fig. 2, GaitMA can be divided into five parts: joint/limb-based heatmap generation, multi-model spatial-temporal encoding, co-attention alignment module, mutual learning module, loss function.

A. Joint/limb-based Heatmap Generation

GCN is operated on an irregular graph of skeletons [5], [6], which makes it difficult to fuse with other modalities usually represented on regular grids. we represent each frame of skeleton points as a joint/limb-based heatmap to improve the effectiveness of modality combination [12], [13]. By creating Gaussian heatmaps centered at each skeleton point using coordinate triplets (x_k, y_k, c_k), we obtain the joint-based heatmap $\mathcal{J}$ with dimensions of $K \times H \times W$, where K represents the number of joints, and H and W denote the height and width of the frame. The formulation is expressed as,

$$\mathcal{J}_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k. \quad (1)$$

The parameter σ regulates the variance of the Gaussian maps, while (x_k, y_k) represents the spatial location of the k-th joint, and c_k represents the corresponding confidence score. We can also create the limb-based heatmap $\mathcal{L}$:

$$\mathcal{L}_{kij} = e^{-\frac{\mathcal{D}((i,j),seg[a_k,b_k])^2}{2*\sigma^2}} * \min\left(\mathcal{C}_{a_k}, \mathcal{C}_{b_k}\right). \quad (2)$$

The limb indexed as k connects two joints, a_k and b_k. The function $\mathcal{D}$ calculates the distance from the point (i, j) to the segment $[(x_{a_k}, y_{a_k}), (x_{b_k}, y_{b_k})]$. Finally, the joint/limb-based

heatmap is derived by stacking all the heatmaps for each frame along the K dimension.

B. Multi-model Spatial-temporal Encoding

To enhance the efficiency of spatial-temporal feature extraction from gait information while minimizing model size, we introduce an innovative asymmetric CNN-based architecture[15] with a dual-branch structure. Opting for a higher resolution of 128x88 in Silhouettes allows for the capture of finer details, whereas a joint/limb-based 64x44 heatmap offers comprehensive spatial shape information with reduced model complexity. The silhouette branch employs a dense 3D-CNN to extract detailed, high-dimensional spatio-temporal features. Concurrently, the skeleton branch supplements feature absent in the silhouette representation, utilizing a streamlined 2D-CNN for spatial feature extraction[16], [17].

This asymmetric approach effectively consolidates robust features from both modalities and efficiently trims the model's parameter count. The silhouette features Y_{sil} and the skeleton features Y_{ske} are extracted from the silhouette feature extractor and skeleton feature extractor, respectively. After that we introduce horizontal mapping [18] and temporal aggregation operations to generate feature representations.

C. Co-attention Alignment Module

Silhouette and skeleton features, inherently distinct modalities, often contain mismatched, redundant, and noisy information, impeding detailed inter-modal interactions. Prior research[16], [17] frequently overlooks this complexity, resorting to basic summation, concatenation, or neglecting information redundancy and noise. Addressing this issue, our proposed Co-attention Attention Model (CAM) leverages a self-attention mechanism to align the feature distributions of these two modalities more closely[19]. This alignment not only facilitates inter-feature interaction but also enhances the overall efficiency of model fitting.

As illustrated in Figure. 2, the input Y_m is obtained by channel-wise concatenating Y_{sil} and Y_{ske}, two fully-connected layers are designed to reduce the number of parameters and achieve the information bottleneck effect. The overall formulation can be expressed as:

$$Y_{score} = \sigma\left(\tau\left(\omega_1 Y_m + b_1\right)\omega_2 + b_2\right), \quad (3)$$

$$Y_{align} = Y_{score} \otimes Y_m + Y_m, \quad (4)$$

ω_1, ω_2, b_1, and b_2 represent the weights and biases of two fully-connected layers, respectively. The symbol τ denotes the ReLU activation function, while σ represents the Sigmoid function. $\otimes$ denotes the element-wise multiplication.

D. Mutual Learning Module

To optimize the utilization of features from both modalities, we introduce the mutual learning module (MLM), leveraging a cross-attention mechanism for the comprehensive fusion of these modal features[19]. While the CAM facilitates interaction between modal features, primarily aiming to harmonize their distributional variances, our MLM extends beyond this by ensuring a thorough integration. We employ a symmetric dual-branch structure, allowing each modality to focus on its intrinsic information while concurrently enriching the other. This approach not only enhances the discrete skeleton representation but also augments the semantic content of the silhouette images, achieving a balanced and in-depth feature interaction between the modalities.

The detailed process is shown in Figure. 2. Take one side for example, assuming that Y_1 and Y_2 are the corresponding feature representations of the two modalities, $Y_1^{'}$ is the output after mutual learning. The formulation is expressed as:

$$Y_1^{'} = \Phi\left(\Theta\left(Y_1 Y_2^T/\sqrt{d}\right)Y_2 + Y_1\right). \quad (5)$$

Φ is the layer normalization and Θ denotes the Softmax function, hyperparameter d denotes the scale factor.

E. Loss Function

To achieve optimal performance, we employ triplet loss [22], cross-entropy loss, and Wasserstein loss [14] to train GaitMA.

First, the network is trained to converge by optimizing the classification space using cross-entropy loss which can be formulated as:

$$\mathcal{L}_{\mathrm{ce}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \quad (6)$$

where x_i is the feature of the i-th sample, and its label is y_i.

Second, triplet loss is proposed to enable the model to find a more discriminative metric space by optimizing distances, which can be defined as:

$$\mathcal{L}_{tri} = \varphi\left[D\left(F_i, F_k\right) - D\left(F_i, F_j\right) + m\right]. \quad (7)$$

φ is equal to $max\left(\alpha, 0\right)$, $D\left(F_i, F_k\right)$ represents the Euclidean distance between the features of sample i and sample k, m denotes the margin for the triplet loss.

Finally, we introduce the Wasserstein loss to minimize the distance between the two modalities, ensuring effective fusion and accelerating the convergence of the model. Assuming that the identity features follow a normal distribution, we can utilize online estimations to calculate the means and covariance matrices of the identity features:

$$\tilde{Y_1} \sim \mathcal{N}(\mu, \Sigma), \tilde{Y_2} \sim \mathcal{N}(\mu^*, \Sigma^*). \quad (8)$$

The similarity between these two Gaussian distributions is measured using the 2-Wasserstein distance, which results in the Wasserstein loss:

$$\mathcal{L}_w \triangleq W_2(\tilde{Y_1}, \tilde{Y_2}) = ||\mu - \mu^*||_2^2 + ||\Sigma^{\frac{1}{2}} - \Sigma^{*\frac{1}{2}}||_F^2. \quad (9)$$

The joint loss function can be expressed as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\mathrm{tri}} + \alpha_2 \mathcal{L}_{\mathrm{ce}} + \alpha_3 \mathcal{L}_{\mathrm{w}}, \quad (10)$$

where the hyper-parameters α_1, α_2 and α_3 are balance factors to weight the losses to each other, where α_1 = 1.0, α_2 = 0.1 and α_3 = 0.1 respectively.

TABLE I
QUANTITATIVE COMPARISON OF GAIT RECOGNITION METHODS ACROSS THREE AUTHORITATIVE DATASETS, INVOLVING OUMVLP, GREW, AND GAIT3D. THE BEST PERFORMANCES ARE IN **BLOD**, THE SECOND BEST METHODS ARE UNDERLINED.

Modality	Method	Testing Datasets								
		Gait3D				OU-MVLP	CASIA-B			
							NM	BG	CL	Mean
		Rank-1	Rank-5	mAP	mINP	Rank-1	Rank-1			
Sihouette	GaitSet(AAAI19)[1]	36.7	59.3	30.0	17.3	87.1	95.0	87.2	70.4	84.2
	GaitPart(CVPR20)[2]	28.2	47.6	21.6	12.4	88.5	96.2	91.5	78.7	88.8
	GaitGL(ICCV21)[20]	29.7	48.5	22.3	13.6	89.7	97.7	94.5	83.6	91.8
	GaitBase(CVPR23)[21]	64.6	-	-	-	90.8	97.6	94.0	77.4	89.7
Skeleton	GaitGraph(ICIP21)[5]	8.6	-	-	-	4.2	86.4	76.5	65.2	76.0
	GaitGraph2(CVPRW22)[6]	11.2	-	-	-	70.6	80.3	71.4	63.8	71.8
	GaitTR(ES23)[7]	7.2	-	-	-	39.7	94.7	89.4	86.7	90.2
	GPGait(ICCV23)[8]	22.4	-	-	-	59.1	93.6	80.2	69.3	81.0
Fusion	BiFusion(MTA23)[16]	-	-	-	-	89.9	**98.7**	96.0	92.1	95.6
	MMGaitFormer(CVPR23)[17]	-	-	-	-	90.1	98.4	96.0	**94.8**	**96.4**
	Ours	**66.1**	**81.2**	**55.4**	**34.7**	**91.2**	98.2	**96.7**	92.8	95.9

TABLE II
THE MEAN RANK-1 ACCURACY (%) ON OUMVLP EXCLUDING THE UNDER DIFFERENT SKELETON REPRESENTATIONS, EXCLUDING IDENTICAL-VIEW CASES.

Skeleton Input	Method	Rank-1
Point	BiFusion[16]	89.9
	MMGaitFormer[17]	90.1
Joint-based heatmap	Ours	90.8
Joint/Limb-based heatmap		**91.2**

III. EXPERIMENTS

A. Datasets

We evaluated our proposed method on three commonly used datasets, including one outdoor dataset: Gait3D [23] and two indoor datasets: CASIA-B [24], OU-MVLP [25].

Gait3D [23] is a large-scale gait dataset captured in the wild, comprising 4,000 subjects and 25,309 sequences. The dataset features 25,309 sequences acquired through camera capture and provides four modalities: silhouettes, 2D and 3D coordinates of joints, and 3D meshes. It is divided into a training set containing 3,000 subjects and a test set consisting of 1,000 subjects.

OU-MVLP [25] contains 10307 subjects, and each subject includes 28 sequences obtained from 14 camera views. For each view, each subject has 2 sequences (NM#01 and NM#02). The sequences of the first 5153 subjects were used for training, and the sequences of the remaining 5154 subjects were used for testing.

CASIA-B [24] is one of the earliest widely used gait datasets, consisting of 124 subjects. Each subject is represented with 11 views, and each view contains ten sequences. These sequences are captured under three different walking conditions: normal walking (NM), walking with a bag (BG), and walking in a coat (CL). The dataset is divided into two parts: the first 74 subjects are designated as the training set, while the remaining 50 subjects constitute the test set.

B. Experimental Settings

For CASIA-B and OU-MVLP, the resolution of silhouettes we take is 64×44. For Gait3D, the resolution of silhouettes we take is 128×88. We use SGD as the optimizer for the training model in both CASIA-B, OU-MVLP, and Gait3D. The initial learning rate and the weight decay of the SGD optimizer as 0.1 and 0.0005. For CASIA-B, we train our model for 60k with (8, 16) batch size, the learning rate is set to 1e-2 at the 20k iteration and 1e-3 at the 40k iteration respectively. For OU-MVLP, the total iteration is 150k with (32, 8) batch size, decaying the learning rate to 1e-2 and 1e-3 at the 50k and 100k iterations. For Gait3D, the batch size is set to (16, 4), the total iteration is 60k.

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

We compare GaitMA to other state-of-the-art (SOTA) gait recognition work, with comparative results detailed in Table I. This comparison encompasses methods based on silhouette-based, skeleton-based, and multimodal three mainstream approaches. Additionally, a focused comparison with other multimodal methods, specifically in terms of skeleton representation, is presented in Table II. These comprehensive evaluation results are sourced from the respective original publications.

Comparison with silhouette-based methods: GaitMA exhibits superior performance on the CASIA-B, OU-MVLP, and Gait3D datasets, underscoring the enhanced gait characterization achieved through the integration of the skeleton feature. This is particularly evident in challenging scenarios, such as the real-world Gait3D dataset and the CASIA-B(CL) dataset, where silhouette quality is compromised by complex backgrounds, occlusions, and camera angles. The incorporation of skeleton features in our method not only demonstrates significant improvements in these conditions but also provides an effective resolution to these challenges. Notably, our method outperforms the current leading GaitBase method by a margin of 1.5%.

Comparison with skeleton-based methods: Our approach surpasses current skeleton-based methods, which are hindered by the limited accuracy of pose estimation algorithms and the

TABLE III
ABLATION STUDY ON THE EFFECTIVENESS OF EACH INDIVIDUAL MODULE ON THE GAIT3D

Methods	Sil	$\mathcal{J}$ and $\mathcal{L}$	CAM	MLM	$\mathcal{L}_{\text{w}}$	Rank-1	mAP
Baseline	✓					59.9	48.9
+ $\mathcal{J}$ and $\mathcal{L}$	✓	✓				64.1	52.8
+ CAM	✓	✓	✓			64.5	53.2
+ MLM	✓	✓	✓	✓		65.3	54.2
+ $\mathcal{L}_{\text{w}}$	✓	✓	✓	✓	✓	66.1	55.4

TABLE IV
UNIVERSALITY STUDY RESULTS ON THE GAIT3D DATASET

Method	Modality	Rank-1
GaitSet[1]	Sihouette	36.7
GaitSet-MA	Silhouette+Skeleton	48.2
GaitPart[2]	Silhouette	28.2
GaitPart-MA	Silhouette+Skeleton	45.8
Ours	Silhouette+Skeleton	66.1

absence of spatial shape features, rendering them less competitive, particularly on large-scale and real-world datasets. Notably, our method achieves a 43.7% higher Rank-1 accuracy on Gait3D compared to GPGait.

Comparison with multi-model methods: Diverging from prevalent multimodal approaches that utilize coordinate points, our method transforms joint points into joint/limb-based heatmaps, enhancing skeleton feature representation. We present a comparison of this method with point-skeleton input methods and different heatmap forms on the OU-MVLP dataset in Table II. Table I outlines our evaluation across the full datasets. Here, we initially apply our multimodal strategy to the real-world Gait3D dataset, subsequently achieving state-of-the-art results on the large-scale OU-MVLP dataset. The CASIA-B dataset, comprising 124 individuals in a simplistic indoor setting, presents a risk of overfitting in large models, which may degrade generalization in test scenarios, we believe that CASIA-B is no longer suitable as a benchmark dataset. Notably, our approach registers an improvement of 1.3% and 1.1% on the OU-MVLP dataset, surpassing BiFusion and MMGaitFormer, respectively.

D. Ablation Study

To validate the efficacy of each component in GaitMA, including joint/limb-based heatmaps, which provides robust skeleton gait features, CAM and MLM for spatial and temporal multi-model feature fusion, Wasserstein loss which makes the distribution of fused features as similar as possible, we conduct ablation studies the Gait3D dataset with results in Table III. Furthermore, we demostrate the universality of our method by applying it to two state-of-the-art silhouette-based gait recognition models, the evaluation results of which are displayed in Table IV.

Ablation Study of joint/limb-based heatmaps. To investigate the impact of incorporating the skeleton branch, which is represented by joint/limb-based heatmaps, we devise a baseline model consisting solely of a single silhouette branch. Remarkably, the inclusion of skeletons leads to a substantial increase in accuracy by **3.8%**, thus performing a significant improvement in the gait recognition task.

Ablation Study of CAM&MLM. The integration of these two modules improves the accuracy by **1.2%** compared to a simple element-wise addition approach. Specifically, CAM yields a **0.4%** improvement, while MLM achieves a **0.8%** improvement, demonstrating the effectiveness of each module. The results highlight that the two modules we proposed effectively facilitate the fusion of two modalities, resulting in a more comprehensive and robust gait representation.

Ablation Study of Wasserstein loss. Wasserstein loss makes the distribution of fused features as similar as possible for each identity. When training GaitMA using Wasserstein loss, the accuracy improved by **0.8%**, demonstrating that the introduction of Wasserstein loss ensures effective fusion and accelerates the convergence of the model.

Universality of GaitMA. we demonstrate the universality of our method by applying it to two state-of-the-art silhouette-based gait recognition models, i.e., GaitSet[1], GaitPart[2]. We denote the models after applying our method as GaitSet-MA and GaitPart-MA. The integrated model incorporates the original structures of GatiSet and GaitPart to encode silhouette features. Then, we introduce joint/limb-based skeleton feature encoding to extract spatial shape information from the skeleton modality and introduce CAM and MLM to realize the fusion of multimodal feature information.

The results on the Gait3D datasets, as presented in Table IV, demonstrate the effectiveness of our proposed method. It exhibits significant improvements in Rank-1 accuracy, with an increase of 36.7 to 48.2 for GaitSet and 28.5 to 45.8 for GaitPart. This consistent enhancement is observed across both models, highlighting the efficacy of our approach.

IV. CONCLUSION

This paper introduces GaitMA, a novel multi-modal gait recognition framework that effectively combines two modalities to obtain a more robust and comprehensive gait representation for recognition. Compared to other multi-modal gait recognition approaches, our method consistently demonstrates superior performance across three mainstream datasets, both indoor and outdoor, and marks the first application of multi-modal methods in the wild. Specifically, the use of Heat-skeletons representations provides clearer structural features of the human body and exhibits enhanced robustness in real scenarios. Furthermore, our well-designed Co-attention alignment module and Mutual learning module, along with the introduction of Wasserstein loss, effectively eliminate redundant features between modalities and integrate efficient gait representations. Our goal is to continue advancing the study of multi-modal feature learning within the field of gait recognition, thereby continuously propelling progress in gait recognition.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China (Grant No. 42106193, 41927805).

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020.

[5] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Gaitgraph: Graph convolutional network for skeleton-based gait recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2314–2318.

[6] T. Teepe, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, "Towards a deeper understanding of skeleton-based gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1569–1577.

[7] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Systems*, vol. 40, no. 6, p. e13244, 2023.

[8] Y. Fu, S. Meng, S. Hou, X. Hu, and Y. Huang, "Gpgait: Generalized pose-based gait recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 595–19 604.

[9] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[10] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[11] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 264–284, 2022.

[12] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2969–2978.

[13] S. Guo, E. Rigall, Y. Ju, and J. Dong, "3d hand pose estimation from monocular rgb with feature interaction module," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5293–5306, 2022.

[14] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," *Advances in neural information processing systems*, vol. 28, 2015.

[15] 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.

[16] Y. Peng, K. Ma, Y. Zhang, and Z. He, "Learning rich features for gait recognition by integrating skeletons and silhouettes," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 7273–7294, 2024.

[17] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial-temporal feature fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 949–17 957.

[18] 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 *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8295–8302.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] 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.

[21] 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.

[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[23] 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.

[24] 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.

[25] 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, pp. 1–14, 2018.