View-embedding GCN for skeleton-based cross-view gait recognition

Md. Zasim Uddin, *Member, IEEE,* Ausrukona Ray, Borsha Das, and Md Atiqur Rahman Ahad, *SMIEEE*

***Abstract*—Gait has emerged as a promising biometric modality due to its non-invasive nature and the ability to capture samples from a distance. Model-based gait recognition using skeleton data conveys rich information that remains invariant to carried objects and clothing variations. However, viewing a person from different angles alters their gait posture, resulting in increased intra-subject variability compared to inter-subject variability. Therefore, we propose a novel framework, View-embedding Modified Residual Graph Convolutional Network (VeMResGCN), for cross-view gait recognition (CVGR) by exploiting two modules: Modified Residual Graph Convolutional Network (MResGCN) and View-embedding Feature Extraction (VeFE) for view-invariant features. A state-of-the-art pose estimation algorithm extracts skeleton key points from raw video input, from which multiple features (e.g., relative joint positions, motion velocities, and bone structures) are computed. The final feature vector for gait recognition is computed by consolidating the features from the MResGCN and VeFE modules. To the best of our knowledge, this work is the first to extract view-invariant features in a unified Graph Convolutional Network (GCN) for skeleton-based CVGR. We evaluate our proposed framework on two of the largest publicly available skeleton datasets, CASIA-B and OUMVLP-Pose, under challenging covariates of clothing variation and carried objects. Results demonstrate that VeMResGCN significantly outperforms state-of-the-art methods with average rank-1 accuracies of 90.3%, 80.7%, and 73.4% for normal, carried object, and clothing variations on CASIA-B, and 71.0% on OU-MVLP in terms of skeleton-based CVGR. These results demonstrate the ability of our proposed framework to maintain superior CVGR performance despite the presence of carried objects and clothing variations. The proposed framework holds strong implications for real-world biometric applications, including robust person re-identification and surveillance systems, where maintaining consistent recognition across varying views and covariates is crucial. The source code will be available on https://github.com/RayAusrukona/VeMResGCN.**

***Index Terms*—Gait recognition, Skeleton, Cross-view, View-embedding, Residual graph convolutional network, Graph convolutional network, Biometrics.**

I. Introduction

Biometrics refers to the recognition of individuals based on their physiological or behavioral traits. Gait recognition, a behavioral biometric modality, involves analyzing an individual's unique walking patterns. Unlike other biometrics, gait recognition does not require active participation from the subject and can be performed unobtrusively at a distance, even when facial or iris features are obscured. This makes it particularly valuable in real-world applications such as social security, crime prevention, and forensic analysis, where cooperation may be limited or subjects are unaware of being observed.

Gait recognition, however, faces several challenges due to its sensitivity to various covariates, such as clothing variations [1], walking speed, occlusions [2], [3], carried objects [4], and camera view variations [5], [6]. These covariates can significantly affect both the gait itself and the features extracted for recognition. Among these, camera view variation is considered one of the most challenging, as changes in viewing angle can lead to significant differences in gait appearance due to self-occlusion, posture changes, and limb movement alterations [5], [7]. In particular, gait features can exhibit substantial intra-subject variation when an individual is viewed from different angles, complicating recognition. The complexity increases further when additional covariates, such as carried objects and clothing variations, are present. In this paper, we address one of the most challenging scenarios: skeleton-based cross-view gait recognition (CVGR) under the simultaneous influence of carried objects and clothing variations, a problem that has a significant impact on real-world applications.

To address these challenges, researchers have explored appearance-based and model-based approaches for gait recognition. Appearance-based approaches often rely on background-subtracted silhouettes, which provide detailed shape and motion information [8]–[11]. However, silhouette extraction is affected by background illumination changes, even if the silhouette extraction algorithm is robust. Moreover, the camera view angles, clothing variations, and the carried objects negatively affect the silhouettes and make them ill-posed. As shown in Fig. 1, carrying a bag or wearing a jacket alters the silhouette's shape, highlighting the limitations of appearance-based approaches.

Model-based approaches, in contrast, reconstruct articulated human models from gait sequences in a kinematic manner, using features such as joint angles, limb lengths, and relative positions [12]. These approaches are less sensitive to appearance variations caused by clothing or carried objects, as they rely on kinematic data rather than appearance. However, they often require high-resolution image sequences, and the process of fitting accurate models can be error-prone, limiting their use in gait recognition. A subset of model-based approaches, known as skeleton-based approaches, has recently emerged as a promising alternative. By leveraging skeleton key points as input, skeleton-based Graph Convolutional Networks (GCNs) have demonstrated significant potential for addressing view variation challenges in gait recognition by focusing on the

MZ Uddin, Ausrukona Ray, and Borsha Das were with Department of Computer Science and Engineering, Begum Rokeya University, Rangpur, Bangladesh. E-mail: zasim@brur.ac.bd

MAR Ahad was with Department of Computer Science and Digital Technologies, University of East London, UK. E-mail: mahad@uel.ac.uk

Manuscript received June 11, 2024; revised ...

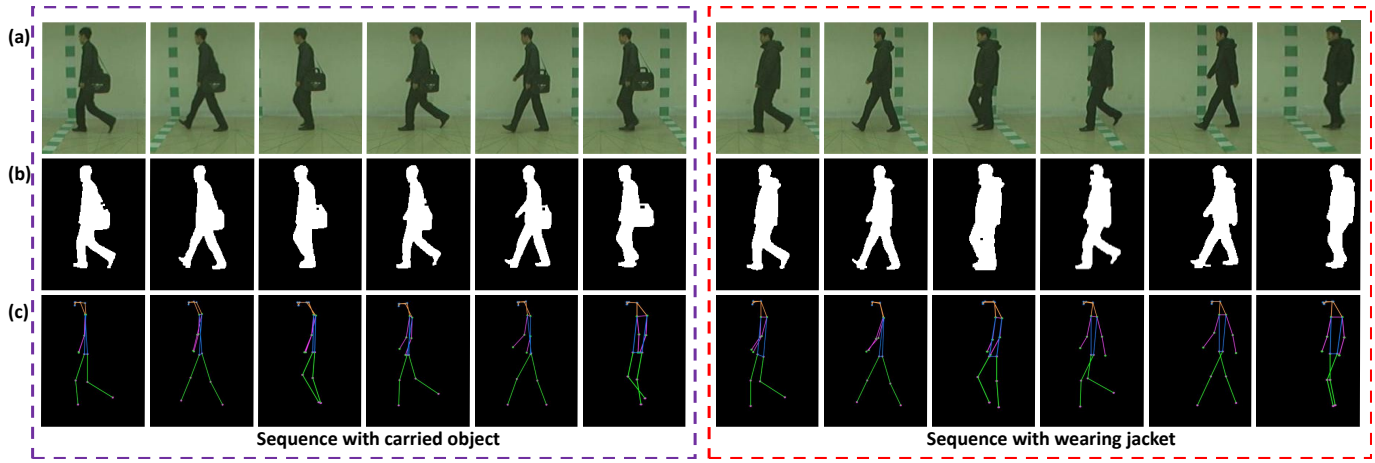

Fig. 1. Example of a gait sequence (every fourth frame) for a subject: (a) RGB image sequence, (b) silhouette image sequence, and (c) skeleton image sequence. On the left, the subject carried an object while wearing a jacket on the right. Carried object and clothing variation negatively affect the corresponding silhouette image shape; however, the skeleton key points are robust against carried object and clothing variations.

extraction of view-invariant features [13]–[17].

Existing skeleton-based GCN approaches, such as ResGCN [13], [14], SDHF-CGN [15], LUGAN-HGC [17], and ResGait [16], perform well when the camera view angle variation between the probe and gallery samples is small. However, they often struggle when large view variations result in shape distortion and self-occlusion. For example, as shown in Fig. 2, front and back view skeleton sequences are prone to self-occlusions, making key points, such as the eyes and ears, difficult to detect. Developing a robust subspace or metric for nonaligned features is still a major challenge, mainly when large view variations and cross-view scenarios are considered. Furthermore, many existing methods do not explicitly detect camera view angles, instead extracting features without adequately accounting for view variability.

This paper addresses the limitations by proposing a novel framework, the View-embedding Modified Residual Graph Convolutional Network (VeMResGCN), for CVGR. Our framework integrates two key modules: Modified Residual Graph Convolutional Network (MResGCN) for discriminant feature extraction and View-embedding Feature Extraction (VeFE) for extracting view-invariant features. The VeFE module explicitly estimates the camera view angle, enabling the framework to extract view-invariant features via a learned projection matrix. The MResGCN module enhances the discriminative power of these features by leveraging residual graph convolutions. Together, these modules improve the accuracy of gait recognition across varying view angles, even in the presence of challenging covariates such as carried objects and clothing variations.

The framework was evaluated on two popular publicly available cross-view gait datasets: CASIA-B [1] and OUMVLP-Pose [18], achieving state-of-the-art accuracy in skeleton-based CVGR. Its robustness against challenging covariates, including carried objects and clothing variations, highlights its potential for real-world security, healthcare, and surveillance applications, advancing the field and supporting practical deployment.

II. Related work

This section explores appearance-based and model-based approaches designed to enhance recognition accuracy and robustness under the challenging covariates such as clothing, carried objects, and view angle variations.

A. Appearance-based approaches

Appearance-based approaches fall into template-based and sequence-based categories. Template-based methods accumulate spatiotemporal gait information into a single template image, with Gait Energy Image (GEI) [8] being a widely used example. GEI aggregates silhouette sequences over one gait cycle, yielding strong performance without covariates [18]. However, real-world systems often encounter covariates, such as carried objects or clothing variations. To address these challenges, subspace and metric learning methods were applied to GEIs, for example, principal component analysis and linear discriminant analysis in [8], as well as RankSVM employed in [4], [19]. More recently, Convolutional Neural Networks (CNNs) have been used to extract discriminative features from GEIs. Shiraga et al. [20] introduced a simple CNN using a single GEI as input with cross-entropy loss, while Siamese networks, paired with contrastive loss, have been employed in [4], [18]. However, finding a robust subspace or metric remains challenging, particularly under large view differences where human gait shapes become highly distorted, and motion information is averaged in template-based methods.

Recently, researchers explored a silhouette sequence-based approach as input for gait recognition. For example, Chao et al. [11] proposed a concise, effective, and view-invariant model, i.e., GaitSet, which treats gait as a set rather than continuous silhouettes and extracts spatiotemporal features using temporal pooling. Moreover, Fan et al. [10] introduced a part-based model called GaitPart, where part-level spatial and short-term temporal fine-grained motion features were extracted using the frame-level part feature extractor and micromotion capture module, respectively, while GaitGL [9]

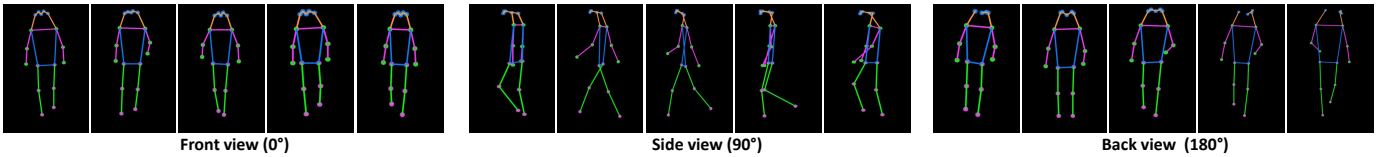

Fig. 2. Skeleton image sequences (every fourth frame) of a subject with multiple views: front, side, and back views.

introduced a fusion framework consisting of a global and local feature extractor followed by local temporal aggregation to extract more detailed local information. While these methods primarily focused on vertical part-based feature extraction, Uddin et al. [6] extended this approach by introducing a Horizontal-Vertical Part Model (HVPM). The HVPM incorporates features extracted along both transverse and sagittal planes, significantly enhancing CVGR by providing a more comprehensive representation of spatial and temporal gait patterns.

In addition, some methods employ view transformation techniques to align probe and gallery features to the same viewpoint [5]. Others incorporate view-dependent features or embed view information during extraction. For example, Chai et al. [7] used a selective projection layer to integrate view information into a state-of-the-art feature extraction framework. Similarly, studies in [21] embedded view information along with the appearance and intrinsic motion information, leveraging Lagrange's equation and a second-order motion extraction module to mitigate viewpoint diversity and intra-class variations.

B. Model-based approaches

Traditional model-based approaches: In the early era of gait recognition, several methods were developed using manually modeled human body shape and motion during walking [12], [22], [23]. Particularly, gait features were extracted from the key joints and/or body parts, such as the position of the hips, knees, ankles, and feet. For example, studies in [23] used a stick or pole to construct the model, where several body points were extracted from the gait silhouette to generate a stick figure for gait recognition, whereas approach in [22] developed a markerless automatic human gait identification system that used a consecutive series of planar 2D sticks to represent gait motion. However, the problem with these model-based approaches is that they require high-resolution images; otherwise, the joint position and joint angle, due to inaccurate point estimation, produce inferior performances.

Skeleton-based approaches: The advent of DL-based pose estimators like OpenPose [24], AlphaPose [25], and HRNet [26] has revolutionized model-based approaches. These algorithms extract skeleton key points from RGB images, providing joint locations along with estimated detection confidence scores. The robustness of skeleton data against traditional covariates [15], such as clothing variations and carried objects, has catalyzed a shift from silhouette-based to skeleton-based gait analysis. Fig. 1 illustrates this advantage: while silhouettes are distorted by clothing and carried objects, skeleton motion primarily reflects the subject's movement dynamics. This invariance to appearance changes, coupled with the rich spatiotemporal information in joint trajectories, has positioned skeleton-based methods at the forefront of gait recognition research.

Some studies introduced DL-based approaches using skeleton data for gait recognition in the literature. For example, Liao et al. [27] proposed PoseGait, which exploits spatiotemporal gait features using CNN with human prior knowledge. They used a combination of features from pose, angle, limb, and motion based on 3D information. On the other hand, transformer-based approaches were explored for skeleton-based gait recognition. For example, Li et al. [28] proposed an autoencoder-based method to disentangle view, motion, and body features explicitly that can be reconstructed from a different view. They obtained aggregated gait features by CNN. Moreover, Zhang et al. [29] introduced a method called Gait-TR, which used a spatial transformer to extract gait features with high accuracy and more robustness, whereas the approach in [30] introduced a heterogeneous spatiotemporal axial mixer to learn the discriminative gait feature with multifrequency signals effectively. It performs a spatial self-attention mixer followed by a temporal large-kernel convolution mixer.

Graph Convolutional Network-based approaches:

The success of skeleton-based Graph Convolutional Networks (GCNs) in action recognition, such as spatial-temporal graph convolutional networks (ST-GCN) [31] and Residual GCN (ResGCN) [32]), has inspired their exploitation to gait recognition, particularly for tackling the view variation problem. A direct approach to tackle skeleton-based CVGR is to extract robust features from a sample of a query (probe) subject and match the corresponding feature of the sample for the subject of the gallery regardless of the viewing angle from which the gait is observed. This approach is known as view-invariant gait recognition. Following the concepts of skeleton-based GCNs for action recognition, the researcher used the GCN for gait recognition. For example, Liu et al. [15] explored the symmetry of human walking, such as the relationship between the left and right legs and hands, to capture the dependencies in dynamic motion from skeleton data. These approaches worked better than the traditional model-based approaches.

Furthermore, following the proven practicability of ResGCN in action recognition, some studies explored ResGCN architecture for gait recognition. For example, the approaches in [13], [14] explored ResGCN architecture for gait recognition using 2D skeleton data; they fuse the multiple features from skeletons, such as bones, velocities, and joints. Similarly, Gao et al. [16] extracted features from the skeleton sequence to obtain spatiotemporal dynamics using the same architec-

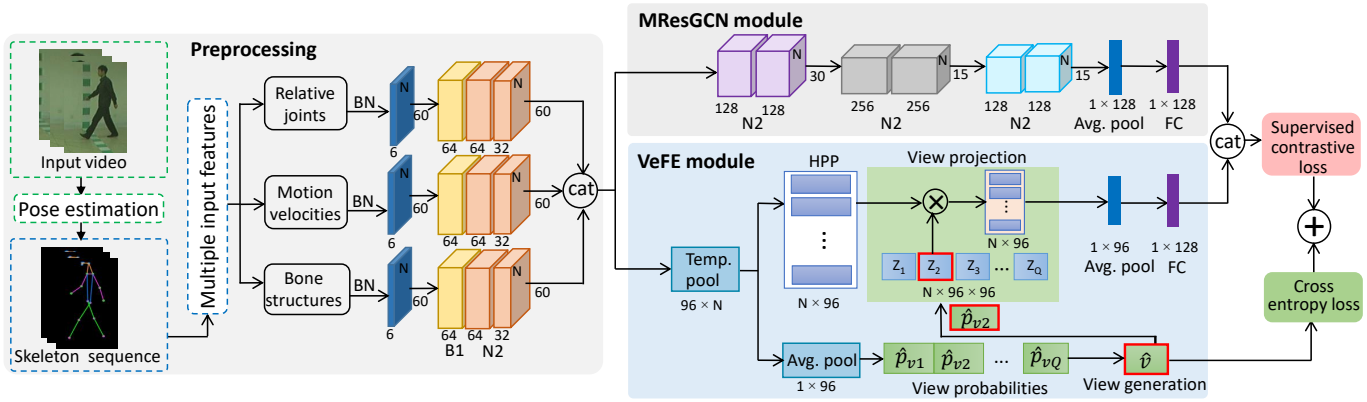

Fig. 3. Overview of the proposed gait recognition framework (VeMResGCN): The left side of the framework is for skeleton preprocessing, the top right module is MResGCN, and the bottom right module is the VeFE. Given a raw RGB video sequence, the skeleton key points were extracted using pose estimation algorithm, and skeleton key points were preprocessed to generate multiple input features. The MResGCN and VeFE modules are then extracted features and combined to make the final feature for gait recognition. The FC, cat, $\otimes$, and $\oplus$ indicate the fully connected layer, concatenation, element-wise multiplication, and addition, respectively.

ture and used the thresholding technique to eliminate noise-related redundant features due to covariate conditions, while Ray et al. [33] introduced a multi-biometric framework that combines features from multiple pose estimation algorithms, utilizing both feature- and decision-level fusion to enhance gait recognition accuracy. In addition, some studies explored the attention module to the GCN. For example, Wang et al. [34] introduced MS-Gait, including multistream architecture along with channel-wise attention to GCN, while Fu et al. [35] presented a part-aware GCN for efficient graph partition and local-global spatial feature extraction. Moreover, similar to the conventional generative adversarial network to image or video, Pan et al. [17] presented a geometry-based multiview pose generation pipeline named lower-upper GAN (LUGAN), where the generator learns a full-rank transformation matrix from the source pose sequence to target view. Particularly, these studies concentrate on extracting apparent information and fusing spatial or temporal features. However, a robust subspace or metric for nonaligned features is difficult to find, especially when the view differences are large due to shape distortion. Our work directly addresses this limitation. In this paper, we propose a unified framework including the view-embedding feature extraction (VeFE) to overcome the limitations of the view angle difference between the samples of probe and gallery for CVGR. Moreover, we introduce modified Residual GCN (MResGCN) for discriminative feature extraction.

III. Proposed Method

A. Overview

We propose a View-embedding Modified Residual Graph Convolutional Network (VeMResGCN) for skeleton-based CVGR. The overall framework is illustrated in Fig. 3. Initially, a raw video sequence is taken as input. Then, a state-of-the-art pose estimation algorithm is used to estimate skeleton key points along with an estimation confidence score to extract multibranch features, including relative joint positions, motion velocities, and bone structures. Within our proposed framework, we have two pivotal modules: (i) Modified Residual Graph Convolutional Network (MResGCN) and (ii) View-embedding Feature Extraction (VeFE). The MResGCN is applied to capture discriminant spatiotemporal features for gait recognition along with the view-invariant feature using the VeFE module. Eventually, the extracted features from the MResGCN and VeFE modules are aggregated to make final features for gait recognition. The feature extraction and gait recognition are trained in an end-to-end manner.

Preliminaries: In accordance with the work of MResGCN, we first construct the skeleton data into a graph to provide input to the model. The skeleton graph is denoted by $G = (V, E)$, where $V = \{v_1, ..., v_N\}$ is the set of N number of nodes representing the joints, and E is the set of edges representing the bones captured by an adjacency matrix $A \in \mathbb{R}^{N \times N}$, which is an undirected graph. If v_i and v_j are connected by an edge, then $A_{i,j} = 1$; otherwise, it is 0. Thus, gait can be considered a sequence of graphs, which has a set of feature tensor $X = \{x_{t,n} \in \mathbb{R}^C \mid t, n \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N\}$ in a temporal dimension T, where $x_{t,n} = X_{t,n}$ is the C dimensional feature vector for node v_n at time t over T, and C is a tuple of 2D coordinate (x, y) and pose estimation confidence score c. Hence, the input gait sequence can be described structurally by the adjacency matrix A and the pose feature tensor $X \in \mathbb{R}^{T \times N \times C}$.

Graph convolutional network: Following the GCN proposed in [36], we used the multilayer GCN with layer-wise propagation rule at time t, denoted as follows:

$$X_t^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}X_t^{(l)}W^{(l)}\right) \tag{1}$$

where $\tilde{A} = A + I_N$ is an adjacency matrix of the gait graph G with identity matrix I_N to obtain the value of the self node. The $\tilde{D}$ is a diagonal degree matrix of $\tilde{A}$, and $W^{(l)}$ is a trainable weight matrix for l^{th} layer. Here, $\sigma(\cdot)$ denotes an activation function, such as the $ReLU(\cdot) = \max(0, \cdot)$, and $X_t^{(l)}$ is the matrix of activations in the l^{th} layer.

B. Human pose estimation and preprocessing

This entails the prediction of the human skeleton and its subsequent preprocessing to prepare the input data for the network. Human pose estimation predicts the spatial locations of human body joints from an image of a gait video sequence. These body joints typically include the eyes, ears, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles and contain the spatiotemporal discriminating gait information for each subject. The task is to estimate the 2D coordinates of N body joints, represented as $X = (x_1, y_1, c_1), (x_2, y_2, c_1), ..., (x_N, y_N, c_N)$, where (x_i, y_i) denotes the 2D spatial position of the i^{th} joint, and N is the total number of joints, whereas c is for confidence s core of each joint estimation. Deep learning-based pose estimation algorithms are classified i nto t op-down a nd b ottom-up approaches: the top-down approach detects the human first and then estimates the body parts, whereas the bottom-up approach detects all body parts first and then groups the parts belonging to distinct persons. Top-down methods use global contexts and structural correlation, and the performance of top-down models is related to human detection results. In this study, we considered the state-of-the-art top-down and bottom-up approaches for human pose estimation, such as OpenPose [24] for bottom-up and AlphaPose [25] and HRNet [26] for top-down, to extract skeleton key points.

Data preprocessing is an important step for skeleton-based gait recognition; according to previous studies for action and gait recognition [14], [32], [37], we exploited the multiple data preprocessing techniques from the raw skeleton data points, including (i) relative joint positions (RJP), (ii) motion velocities (MV), and (iii) bone structure (BS) features. Suppose that the original skeleton key points for a gait sequence are $X = \left\{x \in \mathbb{R}^{T\times N\times C}\right\}$, where T, N, and C denote the number of frames, joints, and 2D coordinates with an estimated confidence score, respectively. The RJP is basically the skeleton key point location compared with the center joint cj of the skeleton. It can be calculated as $RJP = \{r_i | i = 1, 2...., N\}$, where r_i is the concatenation of $x\,[:, i, :] - x\,[:, cj, :]$ and $x\,[:, i, :]$.

Similarly, the second type of data preprocessing is motion velocities (MV), which refers to the changes of joint positions in the next two frames in the temporal dimension. Specifically, it can be a concatenation of $MV_2 = \{f_t | t = 1, 2...., T\}$ and $MV_1 = \{s_t | t = 1, 2...., T\}$, where $MV_2 = \{x\,[t + 2, :, :] - x\,[t, :, :]\}$ and $MV_1 = \{x\,[t + 1, :, :] - x\,[t, :, :]\}$.

The final skeleton preprocessing is bone structure (BS), including bone lengths (BL) and bone angles (BA). To obtain these two sets, the displacement of each bone is calculated as $BL = \{l_i | i = 1, 2...., N\}$, and $BA = \{a_i | i = 1, 2...., N\}$, where $l_i = \{x\,[:, i, :] - x\,[:, i_{adj}, :]\}$, with i_{adj} is the adjacent joint of i^{th} joint, and the angle for each bone is calculated by $a_{i,w} = \left\{arccos\left(\frac{l_{i,w}}{\sqrt{l_{i,x}^2 + l_{i,y}^2}}\right)\right\}$, where $w \in \{x, y\}$ denotes the 2D coordinates.

TABLE I
NETWORK ARCHITECTURE OF THE MODIFIED RESIDUAL GRAPH CONVOLUTIONAL NETWORK (MRESGCN) MODULE.

	CASIA-B			OUMVLP-Pose		
	Block	Layer	Output dimensions	Block	Layer	Output dimensions
MResGCN	Block 0	Batch Norm.	60 × 17 × 6	Block 0	Batch Norm.	30 × 18 × 6
	Block 1	Basic	60 × 17 × 64	Block 1	Basic	30 × 18 × 64
		Bottleneck	60 × 17 × 64		Bottleneck	30 × 18 × 64
		Bottleneck	60 × 17 × 32		Bottleneck	30 × 18 × 32
	Block 2	Bottleneck	30 × 17 × 128	Block 2	Bottleneck	15 × 18 × 128
		Bottleneck	30 × 17 × 128		Bottleneck	15 × 18 × 128
		Bottleneck	15 × 17 × 256		Bottleneck	15 × 18 × 256
		Bottleneck	15 × 17 × 256		Bottleneck	8 × 18 × 256
					Bottleneck	8 × 18 × 256
					Bottleneck	8 × 18 × 256
	Block 3	Bottleneck	15 × 17 × 128	Block 3	Avg. pool 2D	1 × 256
		Bottleneck	15 × 17 × 128		FC	1 × 128
	Block 4	Avg. pool 2D	1 × 128	-	-	-
		FC	1 × 128		-	-

C. MResGCN-based feature extraction

We modified the ResGCN architecture [32], which is developed on the basis of the ST-GCN or basic block [31] and bottleneck block. Each basic block consists of sequential execution of spatial graph convolution and temporal 2D convolution, followed by batch normalization and ReLU activation. By contrast, the bottleneck block is introduced based on the concept of a subtle block structure of ResNet [38], which used two 1 × 1 convolutional layers before and after the common convolution layer of the ST-GCN block, respectively. Particularly, the bottleneck block reduces the number of feature channels and parameters, allowing the model training to become faster. A residual link also connects the features before and after each spatial and temporal block to accelerate the model optimization and reduce the learning difficulties.

Our network is based on the ResGCN architecture and consists of hyperparameters of $[B1, N2, N2, N2, N2]$ in a sequential manner, followed by an adaptive average pool, where $B1$ denotes the basic block, while $N2$ for the two ResGCN modules with bottleneck blocks. The extracted feature using the Modified Residual Graph Convolutional Network (MResGCN) can be considered $f_{MResGCN}$, and the overall network architecture is shown in Table I and Fig. 3.

D. View-embedding feature extraction

Inspired by the success of view-embedding into the appearance-based approach [7], [21], in this study, we proposed the use of the VeFE module for skeleton-based CVGR. The concatenating feature map X_{cat} using the preprocessed three different input skeleton features of the relative joint positions (RJP), motion velocities (MV), and bone structure (BS) is obtained after using the first basic ($B1$) and two consecutive bottleneck blocks ($N2$), and fed it as the input for the VeFE module (as shown in Fig. 3).

In accordance with the temporal max pooling using Eq. 2 on the concatenated feature (X_{cat}) in the time direction, the adaptive average pooling function and fully connected operations are performed to extract the view feature f_{view} in Eq. 3.

$$X_{temp} = P_{tem}(X_{cat}) \tag{2}$$

$$f_{view} = FC(P_{AP}(X_{temp})) \tag{3}$$

TABLE II
NETWORK ARCHITECTURE OF THE VEFE MODULE.

		CASIA-B			OUMVLP-Pose	
		Layer	Output dimensions		Layer	Output dimensions
VeFE	Tem. pool (96 × 17)	Avg. pool 1D	96 × 1	Tem. pool (96 × 18)	Avg. pool 1D	96 × 1
		View probabilities	11		View probabilities	14
		View generation	1		View generation	1
		HPP	96 × 17		HPP	96 × 18
		View projection	17 × 96		View projection	18 × 96
		Avg. pool 1D	96 × 1		Avg. pool 1D	96 × 1
		FC	128		FC	128

where P_{tem} denotes the temporal max pooling, and P_{AP}, FC denotes the adaptive average pool and the fully connected layer, respectively.

Next, on the basis of the extracted view feature, the discrete view angle probabilities are estimated as $\hat{P}_v \in \mathbb{R}^Q$. Finally, a discrete view angle with the maximum probability is calculated as follows:

$$\hat{P}_v = W_v f_{view} + B_v \tag{4}$$

and

$$\hat{v} = \arg\max_i \hat{P}_v(i) \tag{5}$$

where Q is the number of discrete views, W_v, B_v are the weight matrices and bias terms, respectively. The predicted view angle is used to generate a robust view-invariant feature for skeleton-based CVGR. A projection matrix $Z_{\hat{v}}| = \{Z_1, Z_2, Z_3, ..., Z_Q\}$ corresponding to the predicted view $\hat{v}$ is trained, where $Z_i \in R^{D\times D}$.

After estimating the camera view angle, the Horizontal Pyramid Pooling (HPP) with N number of skeleton coordinate points is applied on X_{temp} to obtain the feature tensor $f_{HPP} \in \mathbb{R}^{N\times D}$, where D is the output feature dimension. Then, each feature of the f_{HPP} and the corresponding view projection matrix are multiplied for the estimated view angle. After performing matrix multiplication, we apply adaptive average pooling, followed by a fully connected layer, and obtain the view-invariant feature f_{VeFE}. The overall architecture is shown in Table II.

E. Feature aggregation

Finally, we aggregate the output features from MResGCN and VeFE modules, namely, $f_{MResGCN}$ and f_{VeFE} as:

$$f_{final} = cat\begin{Bmatrix} f_{MResGCN} \\ f_{VeFE} \end{Bmatrix} \tag{6}$$

which is used for CVGR.

F. Loss function

We use a combined loss function consisting of Supervised Contrastive Loss (SCL) [39] and Cross-Entropy Loss (CEL) to train the proposed gait recognition framework effectively. The SCL works with all positive and negative pairs in the batch so that samples of the same class are pulled together in the feature space, whereas the samples from different classes are pushed apart, while CEL is used for the classification of the view angles. During training, the final aggregated feature f_{final} is fed into SCL, and the predicted view $\hat{v}$ is provided into CEL to calculate the losses, and the combined loss is calculated as:

$$L_{final} = L_{SCL} + \lambda L_{CEL} \tag{7}$$

where λ is a weighting parameter to control the trade-off between supervised contrastive loss L_{SCL} and cross-entropy loss L_{CEL}. For a random sample $i \in I \equiv \{1...2K\}$ with a batch size of K, L_{SCL} is defined as,

$$L_{SCL} = \sum_{i\in I} \frac{-1}{|Pos\,(i)|} \sum_{m\in Pos(i)} \log \frac{e^{(z_i\cdot z_m/\gamma)}}{\sum_{a\in R(i)} e^{(z_i\cdot z_a/\gamma)}} \tag{8}$$

where z_i and z_m denote the anchor feature and the corresponding positive features of the same subject, $Pos(i) \equiv \{m \in R(i) : y_m = y_i\}$ is the set of indices of all positives in the batch distinct from i, y_i is the label of the i^{th} sample in the batch, z_a is the other sample whose label is different from z_i or z_m, $R(i) \equiv I\backslash i$ is the set excluding i from all data, $|Pos\,(i)|$ is the number of samples in $Pos(i)$, $|Pos(i)|$ is its cardinality, and γ is a scalar temperature parameter.

We used the cross-entropy loss L_{CEL} for view prediction, and it can be defined as,

$$L_{CEL} = -\sum_{j=1}^{K}\sum_{i=1}^{M} y_j \log(p_{ji}) \tag{9}$$

where K is the number of all gait skeleton sequences, and y_j is the discrete ground truth of view of the j^{th} sequence.

IV. EXPERIMENTS

We conducted experiments using two common publicly available benchmark datasets for CVGR, i.e., CASIA-B [40] and OUMVLP [18], to evaluate the performance of our proposed framework. Furthermore, we performed exhaustive ablation studies to verify the effectiveness of the proposed framework components.

A. Datasets and evaluation protocols

CASIA-B dataset [40] is a popular public gait dataset widely used for CVGR including 124 subjects. Each subject has ten sequences with 11 distinct camera view angles for 18° intervals (0°, 18°, . . . , 180°). More specifically, six are for normal walking (NM), two for walking with a carried object (BG), and the remaining two are for clothing variation (CL). We used the HRNet pose estimation algorithm [26] to estimate 2D pose sequences, similar to studies in [13], [14]. For a fair comparison, we strictly followed the popular protocol used in previous studies [13]–[15], [17], where the first 74 subjects are grouped into the training set, and the remaining 50 subjects are a part of the testing set. In the test set, the gallery set contains the first four sequences of NM condition (NM#1-4), and the probe set retains the remaining sequences (NM#5-6, BG#1-2, CL#1-2); more details about training and test protocol are shown in Table III.

TABLE III
EXPERIMENTAL SETTING ON CASIA-B DATASET

Training	Test	
	Probe	Gallery
Subject ID: 001-074 Seqs.: NM#1-6, BG#1-2, CL#1-2	Subject ID: 075-124 Seqs.: NM#5-6, BG#1-2, CL#1-2	Subject ID: 075-124 Seqs.: NM#1-4

OUMVLP dataset [18] is the world's largest multiview gait dataset, comprising 10307 subjects. Each subject has 14 views in a 15° interval (0°, 15°, . . . , 90°, 180°, 195°, . . . , 270°), and each view has two sequences (#00, #01). Each sequence comprises 18–35 frames and mostly contains approximately 25 frames. Instead of providing RGB image sequence, they release the skeleton pose sequence for publicly available, OUMVLP-Pose [41], using the OpenPose [24] and AlphaPose [25] pose estimation algorithms. Following the official instructions [41], 5153 subjects were used for training, while the remaining 5154 subjects were used for testing. For testing, #00 and #01 sequences were used as the probe and gallery sets, respectively.

B. Training and test details

Training. For the CASIA-B dataset, the Adam optimizer with a one-cycle learning rate scheduler was used to optimize our proposed framework, and the loss temperature γ and hyperparameter λ were set to 0.01. The sequence length, T, and batch size were set to 60 and 128, respectively. In training step one, the maximum learning rate and weight decay were set to 1e-2 and 1e-5 for the first 300 epochs, while 1e-3 and 1e-6 for the successive 100 epochs. By contrast, sequence length, batch size, and learning rate are set to 30, 768, and 5e-3, respectively, for the OUMVLP-Pose datasets, and Training continued up to 950 epochs. Here, stochastic weight averaging [42] was applied after 80% of the maximum epochs.

In addition, the skeleton sequences were flipped from left to right, and various uniform noises were added to augment the training dataset. Moreover, if the length of the skeleton sequence was less than 60 frames for CASIA-B and 30 frames for OUMVLP, the remaining frames were selected from the start of the original sequence and padded to the end. Each experiment was conducted on a single NVIDIA 3090 GPU with PyTorch.

Test. At testing, Euclidean distance was used to calculate the distance between the feature of samples of the probe and the gallery. The Rank-1 identification rate was used to evaluate the performance of the proposed framework.

C. Comparison with skeleton-based state-of-the-art methods

Evaluation on the CASIA-B dataset. We compared the performance achieved by our proposed framework with results obtained using state-of-the-art model-based gait recognition approaches using skeleton data. More specifically, PTSN [27], PoseGait [27], Siamese [43], Disentanglement [28], GaitGraph [13], GaitGraph2 [14], SDHF-CGN [15], ResGait [16], and LUGAN-HGC [44], where GaitGraph2 [14] represent the baseline of our proposed framework. The Rank-1 accuracy on CASIA-B is presented in Table IV. Our proposed framework exhibited superior performance. For example, with a mean accuracy of Rank-1 under 11 probe views, excluding identical-view cases, our proposed framework achieved the best accuracy for the NM and BG covariate and the second-best accuracy for the CL condition. Regarding the NM and BG covariates, our proposed framework achieved a mean accuracy of Rank-1 under 11 probe views, excluding identical-view cases of 90.3% and 80.7% for NM and BG conditions, respectively. This result indicates that our proposed framework increases by 0.7% and 1.0% compared with the second-best approaches, i.e., ResGait [16] and LUGAN-HGC [17], respectively.

The proposed framework surpasses the baseline approach, GaitGraph2 [14], with improvements of 8.3%, 7.5%, and 9.8% for NM, BG, and CL conditions, respectively. This demonstrates its effectiveness in addressing challenging CVGR tasks, particularly under carried object and clothing variation covariates. Notably, our approach consistently ranks as either the best or second-best across all probe view angles. The VeFE module plays a pivotal role in achieving view-invariance, while the MResGCN module significantly enhances the extraction of discriminative features for CVGR. Although our method performs comparably to LUGAN-HGC [44] in the presence of clothing variations, its superior accuracy in NM and BG conditions highlights the robustness of our framework in real-world scenarios.

Evaluation on the OUMVLP-Pose dataset. The Rank-1 accuracy achieved by our framework on the OUMVLP-Pose [41] dataset is shown in Table V. Using skeleton key points generated by OpenPose [25], our framework demonstrated superior performance over several state-of-the-art methods, including CNN-Pose [43], AGGN [43], Siamese [43], GaitGraph2 [14], SDHF-CGN [15], ResGait [16], and LUGAN-HGC [44]. The average Rank-1 accuracy across all view angles, excluding identical views, reached 50.8%, surpassing SDHF-CGN [15] by 0.5% and GaitGraph2 [14] by 6.9%. These results indicate that the proposed framework effectively extracts view-invariant features under challenging cross-view scenarios.

Furthermore, our framework attained the best or second-best accuracy for most of the separate view angle cases (except for 180°, 195°, and 210° view angles only). However, performance decreased slightly for back-view scenarios, such as 180°, 195°, and 210°, where the OpenPose algorithm struggled to detect certain key points (e.g., the nose and ears). This issue is depicted in Fig. 4, which illustrates the difficulty of extracting accurate skeleton sequences in these challenging angles. Despite these limitations, the proposed framework consistently outperformed existing methods under similar conditions.

By contrast, when AlphaPose [25] was employed to generate skeleton data, the average Rank-1 accuracy increased to 71.0% across all view angles, excluding identical views, representing improvements of 8.8% and 8.0% over SDHF-CGN [15] and GaitGraph2 [14], respectively. These results demonstrate the robustness of the proposed framework, with higher-quality skeleton data significantly enhancing the extraction of discriminative, view-invariant features. Particularly, clear and

TABLE IV
RANK-1 ACCURACY (%) ON CASIA-B DATASET UNDER 11 PROBE VIEWS, EXCLUDING IDENTICAL-VIEW CASES COMPARED WITH OTHER SKELETON-BASED METHODS. VALUES IN BOLD AND ITALIC BOLD INDICATE THE BEST AND SECOND-BEST BENCHMARKS, RESPECTIVELY.

Gallery NM#1-4		0°-180°											
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	mean
NM#5-6	PTSN [27]	34.5	45.6	49.6	51.3	52.7	52.3	53.0	50.8	52.2	48.3	31.4	47.4
	PoseGait [27]	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	Siamese [43]	72.4	81.2	85.6	80.4	79.4	85.0	81.0	77.6	82.5	79.1	80.2	80.4
	Disentanglement [28]	35.1	42.2	47.9	48.4	40.7	42.1	42.6	45.3	44.6	37.3	33.9	41.8
	GaitGraph [13]	85.3	88.5	91	***92.5***	87.2	86.5	88.4	***89.2***	87.9	85.9	81.9	87.7
	GaitGraph2 [14]	78.5	82.9	85.8	85.6	83.1	81.5	84.3	83.2	84.2	81.6	71.8	82.0
	ResGait [16]	85.2	***88.4***	**92.8**	90.3	***93.2***	**90.5**	**91.3**	**89.6**	88.6	**89.7**	***85.8***	***89.6***
	SDHF-GCN [15]	77.3	82.8	85.1	86.0	85.5	85.4	83.7	81.5	80.5	83.9	77.6	82.7
	LUGAN-HGC [44]	**89.3**	88.1	89.0	89.9	87.4	88.7	87.4	88.8	**88.8**	87.0	**87.0**	88.3
	VeMResGCN (ours)	***87.7***	**92.5**	***92.5***	**94.2**	**93.7**	***90.1***	***89.3***	88.1	**88.8**	**90.2**	***85.8***	**90.3**
BG#1-2	PTSN [27]	22.4	29.8	29.6	29.2	32.5	31.5	32.1	31.0	27.3	28.1	18.2	28.3
	PoseGait [27]	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	Siamese [43]	62.5	68.7	69.4	64.8	62.8	67.2	68.3	65.7	60.7	64.1	60.3	65.0
	Disentanglement [28]	24.0	29.9	31.3	33.1	29.7	25.6	27.0	29.1	28.6	28.7	28.3	28.2
	GaitGraph [13]	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitGraph2 [14]	69.9	75.9	78.1	79.3	71.4	71.7	74.3	76.2	73.2	73.4	61.7	73.2
	ResGait [16]	73.5	78.2	79.6	***83.3***	**82.4**	***78.5***	**81.7**	***81.1***	***78.4***	80.3	***74.2***	79.2
	SDHF-CGN [15]	67.5	73.9	73.2	74.3	68.5	68.5	70.5	69.0	62.2	68.7	60.1	68.8
	LUGAN-HGC [44]	**79.4**	***79.5***	***81.6***	82.4	78.1	76.2	78.7	**82.0**	**81.6**	**83.0**	73.6	***79.7***
	VeMResGCN (ours)	***78.3***	**82.6**	**84.4**	**86.9**	***78.8***	**83.0**	**82.8**	77.5	***78.4***	***80.7***	***73.7***	**80.7**
CL#1-2	PTSN [27]	14.2	17.1	17.6	19.3	19.5	20.0	20.1	17.3	16.5	18.1	14.0	17.6
	PoseGait [27]	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	Siamese [43]	57.8	63.2	68.3	64.1	66.0	64.8	67.7	60.2	66.0	68.3	60.3	64.2
	Disentanglement [28]	11.6	13.0	15.2	17.0	16.6	17.7	17.8	20.1	19.5	15.3	14.6	16.2
	GaitGraph [13]	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitGraph2 [14]	57.1	61.1	68.9	66.0	67.8	65.4	68.1	67.2	63.7	63.6	50.4	63.6
	ResGait [16]	64.2	68.3	***74.6***	***75.8***	71.6	***72.4***	69.1	70.8	67.6	70.5	67.1	70.2
	SDHF-GCN [15]	63.4	65.4	66.7	64.8	63.0	66.2	69.1	63.3	61.1	65.9	60.7	64.5
	LUGAN-HGC [44]	**72.8**	**72.3**	69.4	75.2	**77.0**	**79.6**	**80.5**	**78.1**	**76.3**	**74.9**	**72.8**	**75.4**
	VeMResGCN (ours)	***72.7***	***70.6***	**76.9**	**77.5**	***74.1***	72.0	***75.5***	***71.2***	***73.2***	***74.6***	***68.8***	***73.4***

TABLE V
RANK-1 ACCURACY (%) ON OUMVLP-POSE DATASET FOR ALL VIEWING ANGLES, EXCLUDING IDENTICAL-VIEW CASES. VALUES IN BOLD AND ITALIC BOLD INDICATE THE BEST AND SECOND-BEST BENCHMARKS, RESPECTIVELY. '-' INDICATES THAT RESULTS ARE NOT AVAILABLE ON THE RESPECTIVE PAPERS.

Method		0° - 90°						180° - 270°								
		0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	Mean
OpenPose	CNN-Pose [41]	8.2	13.9	18.1	22.4	21.3	18.2	10.9	7.3	13.5	12.0	20.5	17.3	13.7	9.4	14.8
	AGCN [43]	19.5	30.5	35.7	38.7	34.0	36.3	29.2	22.7	28.3	27.1	33.6	29.6	30.1	24.9	30.0
	Siamese [43]	25.2	39.0	45.0	48.1	43.5	44.9	36.2	27.7	35.1	34.5	42.4	38.2	39.8	33.5	38.1
	GaitGraph2 [14]	34.7	45.1	49.0	53.7	54.0	51.1	45.0	24.4	34.8	31.6	45.9	52.7	49.5	42.8	43.9
	ResGait [16]	39.6	**49.3**	56.2	58.1	***57.3***	**59.6**	47.7	35.5	40.2	43.3	47.2	**54.9**	**55.3**	***46.2***	49.3
	SDHF-GCN [15]	29.8	*48.6*	***56.5***	**60.1**	55.1	56.9	**51.7**	***44.2***	**49.6**	**48.2**	**55.6**	50.8	***51.2***	**46.3**	***50.3***
	LUGAN-HGC [44]	**42.6**	47.7	52.1	53.6	50.4	51.6	48.0	**49.6**	***45.8***	***47.5***	47.6	47.0	47.4	40.7	47.9
	VeMResGCN (ours)	***41.6***	**55.8**	**58.6**	**61.7**	**62.1**	***58.5***	***49.4***	36.3	44.2	39.8	***53.8***	**55.2**	50.2	43.6	**50.8**
AlphaPose	CNN-Pose [41]	14.3	22.3	27.2	30.0	28.4	23.4	17.2	7.9	13.6	15.6	25.0	24.1	20.2	16.5	20.4
	AGCN [43]	27.3	39.0	45.3	46.5	41.2	46.0	39.9	26.1	30.7	30.5	39.5	35.3	39.3	34.7	37.2
	Siamese [43]	46.5	60.2	68.0	69.3	60.1	66.2	60.7	42.3	51.8	51.7	62.8	55.1	60.8	56.8	58.0
	GaitGraph2 [14]	***53.7***	60.2	64.9	67.2	***66.9***	68.7	63.5	47.7	***58.5***	53.5	***70.0***	***69.9***	***67.7***	***69.3***	***63.0***
	ResGait [16]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	SDHF-GCN [15]	44.5	***61.5***	***70.1***	***72.7***	63.3	***70.0***	***67.6***	***50.6***	57.5	***57.2***	66.8	59.9	64.8	62.9	62.2
	LUGAN-HGC [44]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	VeMResGCN (ours)	**62.7**	**75.0**	**78.1**	**79.7**	**78.8**	**76.4**	**72.0**	**56.5**	**64.6**	**59.6**	**76.1**	**74.4**	**72.3**	**67.3**	**71.0**

accurate skeleton data are critical for reliably estimating view angles and deriving consistent features. As shown in Fig. 4, OpenPose struggled to generate reliable skeleton key points for back view angles (e.g., 180°, 195°, and 210°), whereas AlphaPose yielded a more precise skeleton. These findings highlight the effectiveness of the proposed framework's modules in consistently improving gait recognition accuracy for CVGR.

V. EVALUATION OF THE PROPOSED FRAMEWORK

The proposed framework comprises several key modules: preprocessing, VeFE, and MResGCN-based feature extraction. We conducted various experiments, including different combinations of modules, to evaluate their individual contributions. Furthermore, we evaluated the framework's accuracy against the baseline model, GaitGraph2 [14], to highlight its superior performance across different covariates, including carried objects and clothing variations, and we also compared its results with state-of-the-art appearance-based approaches.

A. Comparison with the baseline model

Here, we present overall comparisons with the baseline model, GaitGraph2 [14]. The results on CASIA-B [40] and OUMVLP-Pose [41] are reported in Figs. 5 and 6. Our proposed framework, i.e., VeMResGCN, outperforms the baseline method, i.e., GaitGraph2, in every aspect. On the basis of these results, the following conclusions are drawn: the accuracy of our proposed framework improves for cross-view cases but is about the same as the baseline for the same view cases. We think the primary cause for the similar accuracy with a baseline

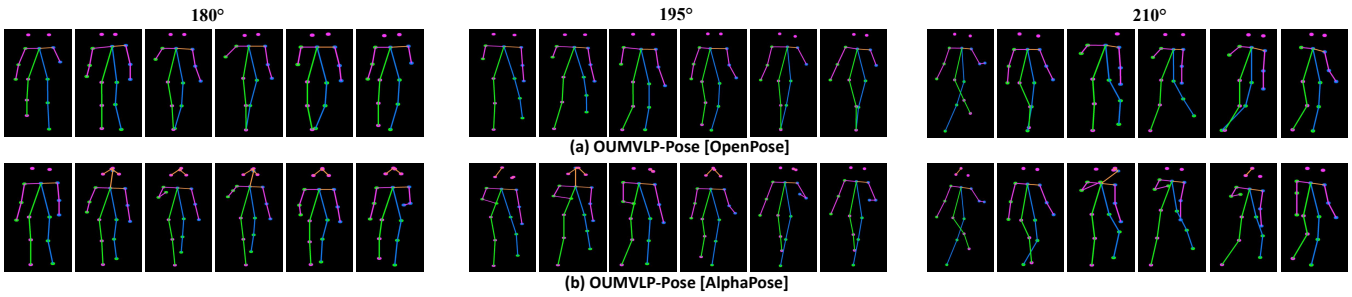

Fig. 4. Example of an extracted skeleton sequence (every fourth frame) for a subject using (a) OpenPose [24]. and (b) AlphaPose [25] pose estimation algorithm for the camera view angle: 180°, 195°, and 210°.

TABLE VI
AVERAGED RANK-1 ACCURACY (%) UNDER THE CONSIDERATION OF DIFFERENT INPUT BRANCHES ON THE CASIA-B DATASET.

Input	NM	BG	CL	Mean
Single input feature	90.4	76.8	70.6	79.4
Multiple input features	90.3	80.7	73.4	81.5

TABLE VII
AVERAGED RANK-1 ACCURACY (%) OF DIFFERENT MODULES ON CASIA-B DATASET.

MResGCN	VeFE	NM	BG	CL	Mean
✗	✗	82.0	73.2	63.6	72.9
✓	✗	84.3	74.5	66.8	75.2
✗	✓	86.4	75.2	69.8	77.1
✓	✓	**90.3**	**80.7**	**73.4**	**81.5**

for the same-view case as probe and gallery samples (e.g., 90° probe angle vs. 90° gallery view angle) is that the VeFE module's ability to estimate view angles and extract view-invariant features is not fully utilized, as the probe and gallery views are essentially aligned. Consequently, the discriminative power of the feature extraction process doesn't significantly improve recognition accuracy since the model doesn't face the challenges of cross-view variation that the VeFE module is designed to overcome.

B. Effectiveness of preprocessing of skeleton data

We used a skeleton preprocessing technique (see Sec. III-B) to extract multiple types of input skeleton features: (i) relative joint positions (RJP), (ii) motion velocities (MV), and (iii) bone structure (BS) features. To analyze the efficiency of the multiple types of input features, we compared accuracy with single input features, such as RJP, similar to GaitGraph [13]. As illustrated in Table VI, the Rank-1 accuracy of our model was 90.3%, 80.7%, and 73.4%, respectively, for NM, BG, and CL conditions when multiple input features are used, while 90.4%, 76.8%, and 70.6% for single-input feature. We can observe that the accuracy is about the same for NM cases as the single-input feature, whereas it improves by 3.9% and 2.8%, respectively, for BG and CL conditions; overall, it surpasses by 2.1%; therefore, multiple-input features demonstrate better in our proposed framework.

C. Effectiveness of VeFE and MResGCN under covariate conditions

The results are presented in Table VII and Fig. 7, showing that our framework not only achieves overall accuracy improvements but also excels in handling the challenging covariate conditions of carried object and clothing variations.

The experimental results highlight the critical role of both VeFE and MResGCN modules in improving recognition performance under challenging conditions. When both VeFE and MResGCN are absent, the framework mirrors the baseline approach and achieves lower accuracy: 82.0%, 73.2%, and 63.6% for NM, BG, and CL conditions, respectively. However, when added only MResGCN improves accuracy slightly for all conditions. For example, accuracy under the clothing variation (CL) increases from 63.6% to 66.8%, showing the module's capability to extract more discriminative features even when subjects' appearance changes significantly. Moreover, added VeFE without MResGCN yields a more pronounced improvement, with accuracy under CL increasing from 63.6% to 69.8%. This result demonstrates that the view-embedding feature extraction is particularly effective at mitigating the challenges posed by carried object and clothing variations, as it produces more stable, view-invariant features. When both modules are combined, the framework achieves the highest performance across all conditions, with an accuracy of 90.3% (NM), 80.7% (BG), and 73.4% (CL). More specifically, the accuracy improved with a greater margin, i.e., 8.3%, 7.5%, and 9.8%, as shown in Table VII. These results highlight the synergy between VeFE and MResGCN, showing that together, they can effectively tackle the covariates.

D. Comparison with appearance-based approaches

The results presented in Table VIII show that appearance-based approaches, such as GaitSet [11], GaitPart [10], and GaitGL [9], outperform our skeleton-based framework in terms of Rank-1 accuracy. GaitGL achieves up to 97.4% for NM and a mean accuracy of 91.8%, while our framework, VeMResGCN, achieves 81.5%. This performance gap is expected since silhouette-based appearance approaches utilize richer image-derived features. Despite this, our skeleton-based framework, which relies solely on skeleton key points, demonstrates competitive results under challenging conditions such as carried objects and clothing variations. Furthermore, skeleton-

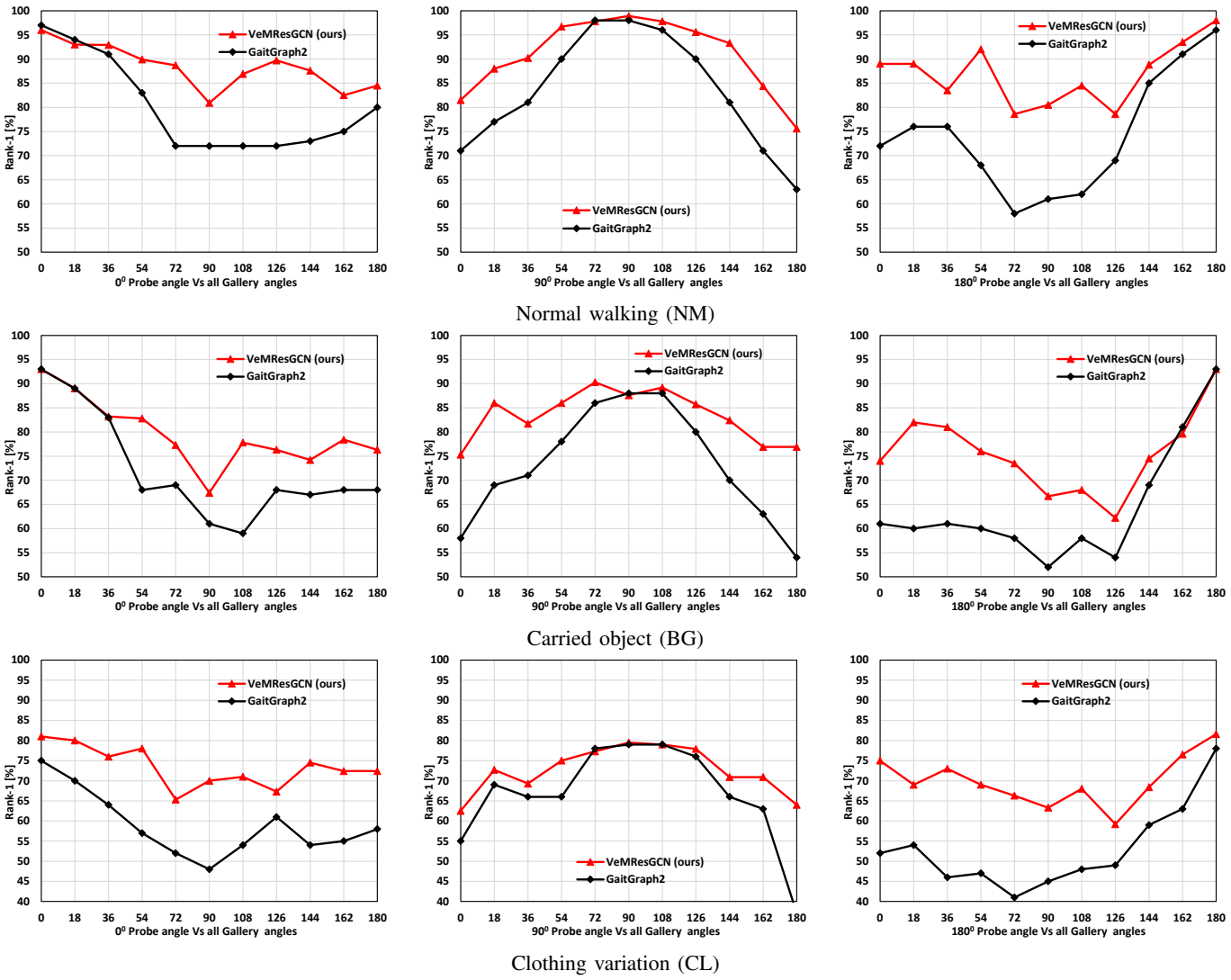

Fig. 5. Comparison of the proposed framework (VeMResGCN) with the baseline method (GaitGraph2) on the CASIA-B dataset. To provide further information, from left to right, three groups of results are respectively shown on three probe view angles, i.e., 0°, 90°, and 180°, against all gallery-view angles. From top to bottom, the result respectively indicates NM, BG, and CL. Best viewed in color.

based approaches offer practical advantages, including low computational cost, robustness to illumination changes, and do not require high-quality images. While these models currently fall short of silhouette-based approaches in accuracy, they present a promising, lightweight solution for real-world gait recognition.

TABLE VIII
AVERAGED RANK-1 ACCURACIES (%) ON CASIA-B: APPEARANCE-BASED VS. SKELETON-BASED APPROACHES.

Type	Method	NM	BG	CL	Mean
Appearance-based	GaitSet [11]	95.0	87.2	70.4	84.2
	GaitPart [10]	96.2	91.5	78.7	88.8
	GaitGL [9]	97.4	94.5	83.6	91.8
Skeleton-based	VeMResGCN (ours)	90.3	80.7	73.4	81.5

VI. CONCLUSION AND FUTURE WORK

This paper presents a framework to address the CVGR problem within a unified approach based on the spatiotemporal patterns of the 2D skeleton sequence. To achieve this, we introduce a novel View-embedding Feature Extraction (VeFE) module combined with Modified Residual Graph Convolutional Networks (MResGCN), which overcomes the limitations of existing methods by explicitly estimating view angles and using them to extract view-invariant features. The proposed VeMResGCN framework leverages the view-invariant features from VeFE alongside the discriminative residual features from the MResGCN module. The aggregated feature enhances the framework's ability to express and discriminate for CVGR. Experimental results on two large-scale cross-view gait datasets demonstrate that the proposed framework achieves superior gait recognition performance using skeleton data, even under different covariates, such as carried objects and clothing variation.

However, the proposed framework's limitation is its dependence on the quality of the underlying pose estimation algorithm. When the pose estimation algorithm fails to accurately detect skeleton key points, particularly in challenging back-view scenarios, recognition accuracy can decrease. Moreover,

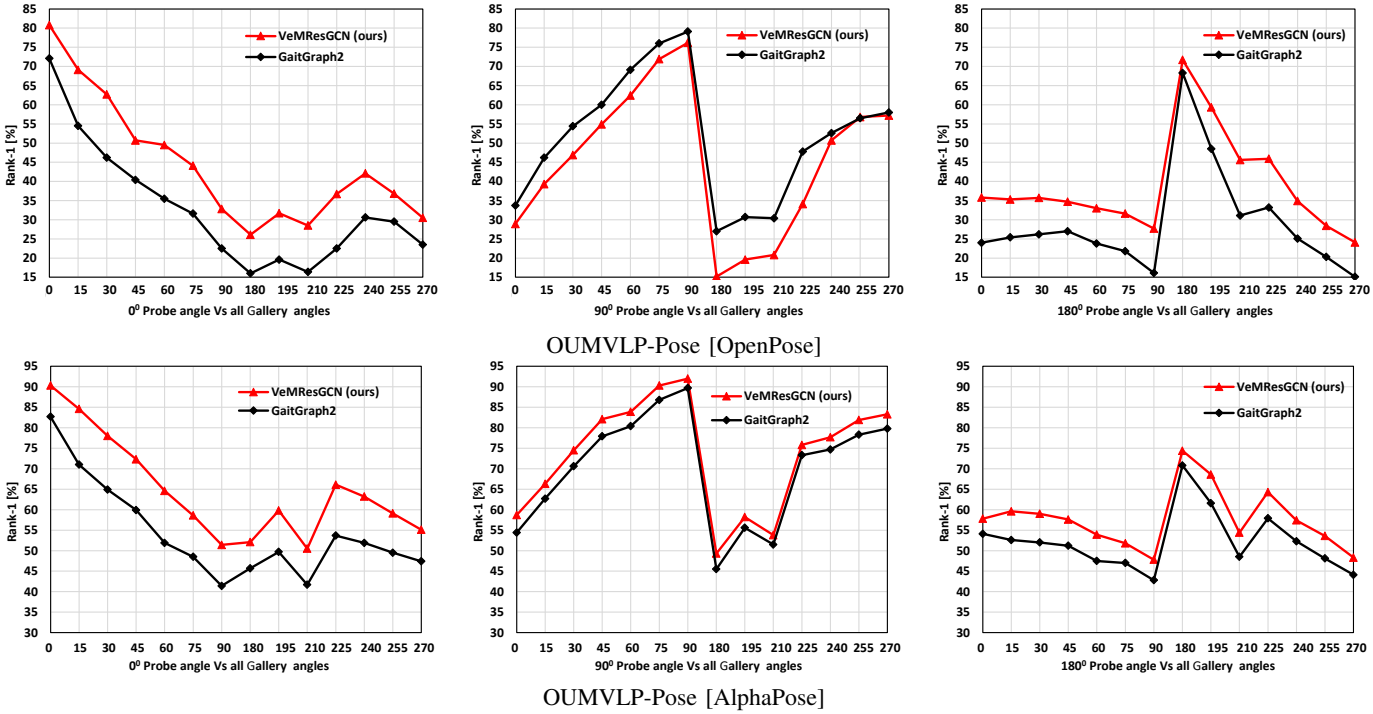

Fig. 6. Comparison of the proposed framework (VeMResGCN) with the baseline method (GaitGraph2) on the OUMVLP-Pose dataset. To provide further information, from left to right, three groups of results are respectively shown on three probe view angles, i.e., 0°, 90°, and 180°, against all gallery-view angles. From top to bottom, the result respectively indicates the skeleton data extracted using OpenPose [24] and AlhpaPose [25] algorithms. Best viewed in color.

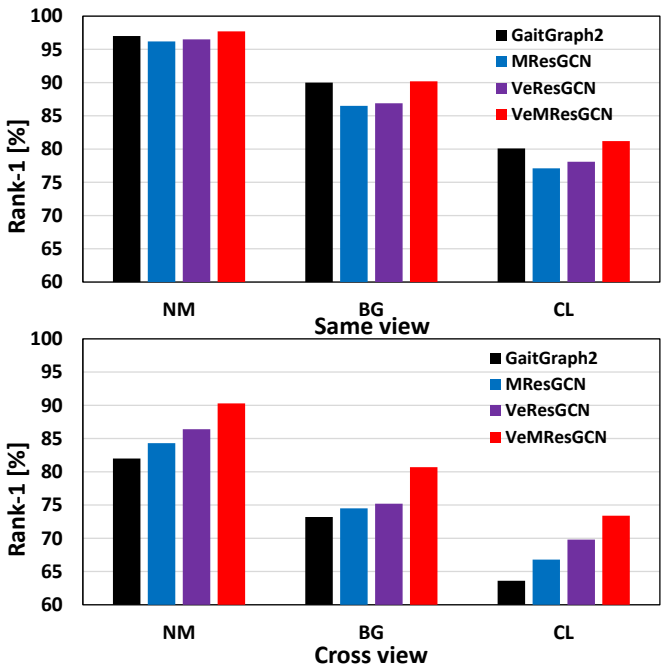

Fig. 7. Average recognition accuracy with different combinations of modules for NM, BG, and CL. Cross-view results are averaged on 11 views, including identical-view cases, whereas same-view results are the average of the same view angle for the probe and gallery. Best viewed in color.

the framework does not show substantial accuracy improvement over the baseline for same-view cases, as the VeFE module's view-invariant feature extraction is less impactful when probe and gallery views are already aligned. This highlights the need for future research to enhance the robustness of skeleton data generation and refine feature extraction under same-view scenarios, thereby minimizing such performance gaps.

We anticipate that the proposed modules will serve as a catalyst for further research in skeleton-based pattern recognition, extending to areas such as view-invariant action recognition and elderly fall detection systems. Furthermore, our VeFE module shows promise for adoption in a variety of applications, including real-time skeleton-based person re-identification.

Acknowledgments

This work was partially supported by the ICT division, Government of the People's Republic of Bangladesh, No.: 1280101-120008431-3631108.

References

[1] 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.

[2] M. Z. Uddin, D. Muramatsu, N. Takemura, M. A. R. Ahad, and Y. Yagi, "Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion," *IPSJ Transactions on Computer Vision and Applications*, vol. 11, no. 1, pp. 1–18, 2019.

[3] K. Hasan, M. Z. Uddin, A. Ray, M. Hasan, F. Alnajjar, and M. A. R. Ahad, "Improving gait recognition through occlusion detection and silhouette sequence reconstruction," *IEEE Access*, 2024.

[4] M. Z. Uddin, T. T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, and Y. Yagi, "The ou-isir large population gait database with real-life carried object and its performance evaluation," *IPSJ Transactions on Computer Vision and Applications*, vol. 10, no. 1, pp. 1–11, 2018.

[5] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Z. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 140–154, 2014.

[6] M. Z. Uddin, K. Hasan, M. A. R. Ahad, and F. Alnajjar, "Horizontal and vertical part-wise feature extraction for cross-view gait recognition," *IEEE Access*, 2024.

[7] T. Chai, X. Mei, A. Li, and Y. Wang, "Silhouette-based view-embeddings for gait recognition under multiple views," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2319–2323.

[8] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 2, pp. 316–322, 2005.

[9] 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.

[10] 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.

[11] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, "Gaitset: Cross-view gait recognition through utilizing gait as a deep set," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3467–3478, 2021.

[12] M. Deng, C. Wang, F. Cheng, and W. Zeng, "Fusion of spatial-temporal and kinematic features for gait recognition with deterministic learning," *Pattern Recognition*, vol. 67, pp. 186–200, 2017.

[13] 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.

[14] 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.

[15] X. Liu, Z. You, Y. He, S. Bi, and J. Wang, "Symmetry-driven hyper feature gcn for skeleton-based gait recognition," *Pattern Recognition*, vol. 125, p. 108520, 2022.

[16] S. Gao, Z. Tan, J. Ning, B. Hou, and L. Li, "Resgait: gait feature refinement based on residual structure for gait recognition," *The Visual Computer*, pp. 1–12, 2023.

[17] H. Pan, Y. Chen, T. Xu, Y. He, and Z. He, "Towards complete-view and high-level pose-based gait recognition," *IEEE Transactions on Information Forensics and Security*, 2023.

[18] 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.

[19] R. Martín-Félez and T. Xiang, "Uncooperative gait recognition by learning to rank," *Pattern Recognition*, vol. 47, no. 12, pp. 3793–3806, 2014.

[20] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *2016 international conference on biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[21] T. Chai, A. Li, S. Zhang, Z. Li, and Y. Wang, "Lagrange motion analysis and view embeddings for improved gait recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 249–20 258.

[22] J.-H. Yoo and M. S. Nixon, "Automated markerless analysis of human gait motion for recognition and classification," *Etri Journal*, vol. 33, no. 2, pp. 259–266, 2011.

[23] G. Ariyanto and M. S. Nixon, "Marionette mass-spring model for 3d gait biometrics," in *2012 5th IAPR International Conference on Biometrics (ICB)*. IEEE, 2012, pp. 354–359.

[24] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[25] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[26] 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.

[27] 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.

[28] Z. Li, S. Yu, E. B. G. Reyes, C. Shan, and Y.-r. Li, "Static and dynamic features analysis from human skeletons for gait recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–7.

[29] C. Zhang, X.-P. Chen, G.-Q. Han, and X.-J. Liu, "Spatial transformer network on skeleton-based gait recognition," *Expert Systems*, p. e13244, 2023.

[30] E. Pinyoanuntapong, A. Ali, P. Wang, M. Lee, and C. Chen, "Gaitmixer: skeleton-based gait representation learning via wide-spectrum multi-axial mixer," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[32] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1625–1633.

[33] A. Ray, M. Z. Uddin, K. Hasan, Z. R. Melody, P. K. Sarker, and M. A. R. Ahad, "Multi-biometric feature extraction from multiple pose estimation algorithms for cross-view gait recognition," *Sensors*, vol. 24, no. 23, p. 7669, 2024.

[34] L. Wang, J. Chen, Z. Chen, Y. Liu, and H. Yang, "Multi-stream part-fused graph convolutional networks for skeleton-based gait recognition," *Connection Science*, vol. 34, no. 1, pp. 652–669, 2022.

[35] 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 (ICCV)*, October 2023, pp. 19 595–19 604.

[36] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[37] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[38] 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.

[39] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[40] 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.

[41] W. An, S. Yu, Y. Makihara, X. Wu, C. Xu, Y. Yu, R. Liao, and Y. Yagi, "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE transactions on biometrics, behavior, and identity science*, vol. 2, no. 4, pp. 421–430, 2020.

[42] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. Association For Uncertainty in Artificial Intelligence (AUAI), 2018, pp. 876–885.

[43] Z. Wang and C. Tang, "Model-based gait recognition using graph network on very large population database," *arXiv preprint arXiv:2112.10305*, 2021.

[44] H. Pan, Y. Chen, T. Xu, Y. He, and Z. He, "Toward complete-view and high-level pose-based gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2104–2118, 2023.