Gait recognition based on 3D point cloud data augmentation

Qingao Yang[1,2,3,4], Xin Chen[1,2,3,4], Yubin Lan[1,2,3,4*], Xiaoling Deng[1,2,3,4*]

(1. *College of Electronic Engineering, College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China*;
2. *National Center for International Collaboration Research on Precision Agricultural Aviation Pesticide Spraying Technology, Guangzhou 510642, China*;
3. *Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, China*;
4. *Guangdong Engineering Technology Research Center of Smart Agriculture, Guangzhou 510642, China*)

Abstract: The goal of gait recognition is to recognize human identity through walking patterns. There are two main methods of gait recognition in the existing research. The first method is based on appearance to extract gait features from binary contour images, and the second method is based on model to extract gait features from key joints. However, the effect of appearance based methods will be affected by the changes of carrying objects and different clothing, while model-based methods will be affected by the effect of pose estimation in recognition, and include sparse gait features, which makes existing gait recognition performances highly depend on visual texture information (such as clothing, carrying and so on). Combining the advantages of the above two methods, we can not only use continuous contour sequences, but also remove the influence of clothing and occlusion. In this paper, we propose a gait feature extraction method based on 3D point cloud. The proposed method first extracts 3D point cloud based on each person's walking video. Two different methods are proposed to map the 3D point cloud data to 2D black and white images. Then the projected 2D images are combined with the original 2D gait samples to expand existing gait datasets. We evaluate 3D point cloud based gait recognition methods on popular gait datasets. The experimental results demonstrate that our proposed method can achieve improvements compared to existing methods, and can achieve the state-of-the-art recognition performances under several experimental settings.

Keywords: Gait recognition, 3D point cloud, 2D mapping, occlusion elimination, dress and carrying variances
DOI: 10.33440/j.ijpaa.20230601.221

Citation: Yang Q G, Chen X, Lan Y B, Deng X L**.** Gait recognition based on 3D point cloud data augmentation. Int J Precis Agric Aviat, 2023; 6(1): 68–82.

1 Introduction

In the era of rapid technological advancement, the synergies between gait recognition technology and precision agriculture aviation have emerged as transformative forces in the agricultural domain. Gait recognition, as a biometric identification technique analyzing individual walking patterns, has showcased versatile applications across various fields. Simultaneously, precision agriculture aviation, leveraging aerial platforms and advanced sensors, facilitates real-time monitoring and data collection, providing intelligent support for agricultural decision-making.

Gait recognition technology enables precise monitoring of personnel activities in agricultural fields. Through the analysis of gait patterns, it becomes possible to identify the location, movement paths, and work efficiency of field workers. This capability empowers agricultural managers to optimize human resources, thereby improving the efficiency of field management. Within the realm of animal husbandry, gait recognition technology proves valuable for monitoring and tracking animal behavior. By analyzing gait characteristics, it becomes feasible to identify the health status, activity patterns, and ecological behaviors of livestock. Real-time data support from gait recognition contributes to informed decision-making in livestock management. The data output from gait recognition can be seamlessly integrated into precision agriculture aviation systems, offering comprehensive and accurate information for agricultural production. Agricultural managers can leverage this information to formulate intelligent decisions, such as optimizing planting schemes, improving irrigation systems, and enhancing crop yields.

By amalgamating gait recognition technology with precision agriculture aviation, the agricultural management landscape experiences notable advancements. This integration not only elevates the level of intelligence in agricultural management but also reduces resource wastage and production costs. The continuous innovation and application of gait recognition technology present myriad possibilities, propelling precision agriculture towards a more intelligent and efficient future. The nuanced insights derived from gait recognition offer novel perspectives and solutions for sustainable development in the agricultural sector.

The goal of gait recognition is to take videos of different people walking, and identify different people through different walking styles[1]. Compared with other recognition technologies,

Received date: 2023-11-15 **Accepted date:** 2023-12-22
Biographies: Biographies: Weixiang Yao, Doctoral student, research interests: precision agriculture aviation technology and equipment, Email: 1913835329@qq.com; **Wesley Clint Hoffmann**, PhD, research interests: agricultural aerial applications. Prology Consulting LLC, College Station, TX 77845, USA, Email: clint.hoffmann@gmail.com; **Shuang Guo**, Postgraduate student, research interests: agriculture aviation technology, Email: 2358681011@qq.com; **Shengde Chen**, PhD, research interests: precision agriculture aviation, Email: 1163145190@qq.com; **Sheng Wen**, PhD, Associate Professor, research interest: precision agricultural aviation application, Email: 58675023@qq.com; **Wen Zeng**, PhD, research interests: precision agriculture aviation, Email: zengwen@scau.edu.cn; **Zhihong Li**, Doctoral student, Technician, research interests: precision agriculture technology and equipment, Email: 402602144@qq.com; **Xiaowen Liang**, Agronomist, research interests: plant protection, Email: xiaowen@jxtianren.com.

*** Corresponding author: Yubin Lan**, PhD, Distinguished Professor, Director, research interests: precision agricultural aviation application, Email: ylan@scau.edu.cn.

gait does not require the active cooperation of people, and the recognition task is completed only by taking walking videos, which greatly reduces the difficulty of obtaining data, and the periodicity of gait has also been proved to be unique. And powerful features, so gait recognition is widely used in various industries in society[2]. However, the task of gait recognition is prone to various interfering factors, leading to unsatisfactory recognition results. Factors such as changes in clothing or the presence of carried objects can obscure the complete gait sequence[3-4]. Therefore, the challenge of gait recognition is to still have a high recognition effect in the presence of negative effects such as realistic occlusion.

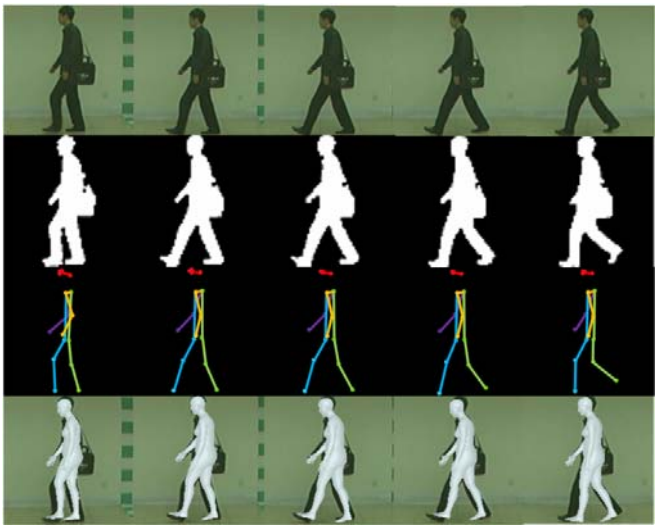

Figure 1 The visual representations of different gait recognition methods. From top to bottom represents the original videos, binary contour images, bone key point features[5] and 3D point cloud features. From the visualizations, we can see that 3D point cloud is quite good at reducing influence of clothes and carryings

In general, gait features can be mainly divided into two types: appearance-based gait features[6-7] and model-based gait features[8-9]. Appearance-based features include more comprehensive features but are easily affected by disturbing factors[10]. In recent years, many existing gait recognition studies use the method of convolutional neural network to extract gait features, which have proved their great recognition ability and recognition effect[11]. Shiraga et al.[12] used 2D neural networks to extract globally valid gait features by taking gait energy images as input. Chao et al.[13] used 2D convolutional neural network extracts global effective features at each frame level, and extracts local effective gait features from local parts of the human body. Zhang et al.[14] divided the human body into different human body parts and used more than one independent 2D neural networks to represent local effective characteristics. Fan et al.[15] proposed a focal convolutional layer method to more effectively extract local gait features for recognition. Global features are not effective enough for the details of human walking gait, while relative local features may pay less attention to the connection between local areas, or even lose context information. Wolf et al. [16] proposed a 3D neural network to extract gait features more effectively, but the traditional 3D neural network requires the length of the gait sequence to be a fixed value, in order to perform classification tasks, and cannot directly model gait features in various-length fragments of gait videos.

However, both appearance-based gait recognition methods and model-based gait recognition methods have limitations to some extent, as shown in Figure 2. Appearance-based gait features are easily affected by changes in perspective, clothing and carrying, while model-based gait features are too sparse to contain sufficient effective information.

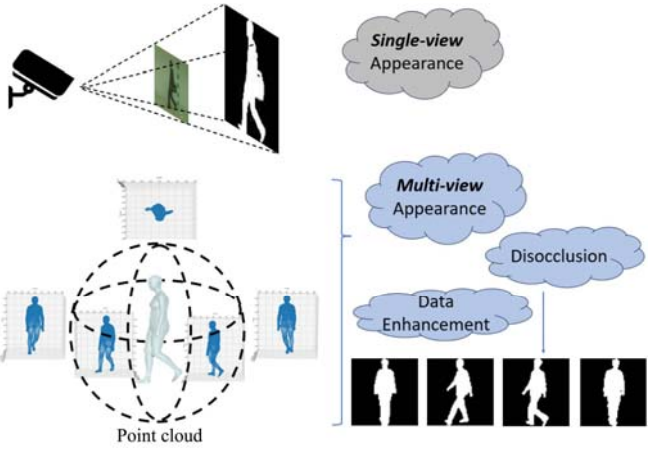

Figure 2 Illustration of Gait Recognition based on Cameras and 3D Pose Estimation. Gait recognition based on cameras typically utilizes silhouettes to learn shape information from a single viewpoint. Gait recognition based on 3D pose estimation leverages three-dimensional structure, shape, and occlusion removal to identify individuals

To address the above issues and combine the advantages of appearance-based and model-based methods, in this paper, we propose to reconstruct three-dimensional (3D) point cloud of human body from single gait images. From the reconstructed 3D point cloud, we can directly extract robust gait features insensitive to texture variances[17-18]. By this way, the influence of dress and carryings will be reduced by a large extent and adequate gait features will be preserved. Then we make full use of the advantages of appearance-based and model-based gait recognition methods.

The 3D point cloud can segment the effective 3D shape information of the human body and the human body texture information. The 3D human body shape information is capable of describing the essential human body structures and robust to dress or carrying variances. Moreover, the 3D point cloud can describe much richer gait information than existing model-based methods using human key joints. 3D human body reconstruction is specifically a single-stage network that predicts multiple 3D characters in a detailed way down to the pixel level. It can predict differentiable images from an image, so as to analyze the 3D mesh information of each person in a simple positive image.

During the experiment, we found that a large number of 3D point cloud features will bring a large computational cost, we need to design a feature extraction method to extract effective gait features from 3D point cloud. In this paper, we design a feature mapping network to map 3D point cloud to 2D gait features. By adding 3D reconstruction loss to the mapping network, the mapping model can decouple 3D shape information from visual texture to 2D view plane, then the docking between 3D and 2D features is established. The mapped 2D features can be used to complete identity recognition.

Traditional binary gait contour images are very sensitive to the occlusion of clothing or carrying conditions, and it is difficult to completely extract the basic features of the human body. Although the method based on key joint points can eliminate the influence of backpack, the extracted joint point information is too limited to reflect complete gait information[19-20]. The 3D point

cloud method can not only eliminate the occlusion effect of clothing or carrying conditions, but also extract adequate gait features. The huge number of accessible gait features in 3D point cloud are undoubtedly capable of improving gait recognition performances.

The main contributions of the proposed methods are as follows:

• We propose to extract gait features based on 3D point cloud method, which makes the huge point cloud features bring more possibilities for high-accuracy gait recognition study.

• We propose a fast-running 2D and 3D double loss function, which realizes the adaptive 2D feature mapping from a large number of 3D point cloud data to 2D features, which reduces the large computational cost caused by huge data amount in 3D point cloud and effectively retains the main gait characteristics in 3D point cloud.

• We propose a fusion method of 3D point cloud features and 2D binary contour features, and the obtained fusion data can more effectively make up the limitations of appearance-based and model-based methods, which is superior to the existing gait recognition methods and presents satisfactory recognition performances.

2 Related Works

2.1 3D mesh regression

3D point cloud reconstruction requires 3D pose estimation. Many research works have transformed the task of 3D human body recognition pose estimation and reconstruction into the task of 3D key points of the human body. These studies are divided into two types, the first one is a single-stage method, which takes an image as input and estimates the 3D key joint points, these methods have impressive effect[21-22] . The second one is a two-stage approach, which first estimates 2D qualifications and then estimates 3D skeleton information in the second stage[23-24] or regression methods[25-27]. The two-stage method requires a pose estimation network with a strong estimation effect, and on this basis, the generated heat map and an effective backbone network are used to comprehensively improve the performance of human recognition.

In subsequent studies, with the development of parametric models, model-based 3D human body shape poses have gradually become more powerful. Compared with modeless methods, parametric human body model can get the human body mesh. With the advancement and development of the field of deep learning, many studies have begun to focus on deep-learning-based methods to improve the performance of 3D human shape pose estimation[28-31]. In 3D pose estimation, the relative rotation position of the shape is difficult to be learned by the network, so many research methods replace this learned feature from the intermediate representation, such as key points and image semantic segmentation[32-34]. There are also some studies [35-36] using advanced deep networks to improve the regression network in the model.

In 3D human body shape and pose estimation, the problem of relative rotation always exists. At the same time, some researches are trying to improve the effect of this aspect. For example, Zhou et al.[37] estimated the different angles of joint rotation in the human body, and Yoshiyasu et al.[38] estimated three-dimensional rotation matrix, Mehta et al. [39] designed a relative rotation program to predict Euler angles to solve the rotation problem identified by 3D estimation.

2.2 Silhouette-based Gait Recognition

Existing popular appearance-based gait recognition methods take gait silhouettes as input and design deep networks [40-43] to extract features. Local gait features of human body parts have been proved to have high discriminative capacity[44-45].

In order to explore more discriminative features embedded in temporal information, some research works directly use the original silhouette sequences as input[46]. Wu et al.[41] designed a 3D convolutional network with multiple adjacent frames as input, and verified in experiments that considering time information can significantly improve performance. Zhang et al.[47] proposed the operation of separating the gait features of primitive people from the original image. Some works[48] created 3D tensors based on the spatio-temporal information available in the sequence, Chao et al.[13] proposed a method that treated gait as a series of silhouettes, instead of isolating gaits. The author designed a series of well-oriented operations according to the cross-view conditions. However, the proposed silhouetted gait sequence is able to incorporate many features not found in previous studies[49], such as changes in walking state, but clothes and carrying objects still have a certain degree of negative effect on the performance of recognition. Fan et al.[15] devised a method to exploit local part features.

2.3 Skeleton-based Gait Recognition

The model-based gait recognition method takes the skeleton key joint points as the input of the network, and the joint point data can be obtained through sensors[50-51] and human body pose estimation methods[52-54]. Early methods[55-56] face difficulties in robust and accurate fitting of human models, so the results are not very satisfactory. On the other hand, recent studies[57-58] overcome these difficulties by using the most advanced human posture estimation methods (such as OpenPose[52] and Human Mesh Restoration (HMR) network[59]), Therefore, it can be found that key points of bones have greater advantages than contours. For example, Liao et al.[57] obtained key point data of human bones through OpenPose, a 2D human body pose estimation network, and then sent the key point data into the network for feature extraction. extraction and identification. Liao et al.[8] conducted a deeper study by extending 2D joint points to 3D, and 3D information is more obvious to view changes. Li et al.[58] used the HRM network to optimize the extraction and recognition network by using a parametric model.

The bone key points can effectively eliminate the influence of the external contours, and are more effective for human gait information. Compared with silhouettes, skeleton dynamics are inherently more robust to identity independent factors (such as wearing and carrying conditions), because skeleton patterns are only concentrated in the essential human body structures. The skeleton information is represented by the position of different joints and the confidence score, which is used to display the detection quality of each point. The most direct method to model these points in the depth neural network is generating the body joint heat maps. Liao et al.[8] CNN is used to obtain spatiotemporal features using human 3D posture. In addition, Since in recent deep learning researches, the application and effect of the graph convolutional neural networks are very good in various fields[60-61], some studies have begun to use the graph convolutional neural network to extract gait features from the key point information of the skeleton. For example, Li et al.[62] first applied GCN to recognize gait, and achieved remarkable performance despite low dimensional features. Teepe et al.[5]

further extract gait feature information effectively from key points of human skeleton based on residual network and graph convolutional neural network[63]. However, these methods are usually more sensitive to occlusion because they rely too much on accurate detection of body joints. Although these works show encouraging improvements, according to human experience, the deep features learned in the neural network model are relevant features, but the basic features that are effective in dynamic mode may be lost in joint position operation. Therefore, the ability of expression and generalization is limited.

3 Method

Through the estimation of 3D body pose and shape, the 3D point cloud features of the human body are obtained. Subsequently, these features are mapped to 2D images using a specially designed autoencoder. To facilitate the training process, both 3D and 2D losses are incorporated as auxiliary losses to guide the network. The resulting images are then combined with the existing binary gait silhouette images, thereby expanding the dataset of gait samples.

3.1 Extracting 3D point cloud

In this section, multiple 3D human pose shapes in an image are estimated by a single-level network in a predictive pixel-level manner in each frame. In the network, multiple differentiable levels of images are predicted from the input image to resolve the 3D pose shape mesh of each person. Specifically, as shown in Figure 3, the 3D human pose shape estimation network predicts two types, the first is the body center heat map representing the position of the 2D human body center, and the second is representing Mesh parameter map of the 3D mesh parameters of the human body. Through such two sampling processes, the center position obtained from the body center heatmap and the 3D person body mesh parameter map are combined to put the parameters into a parametric body model[64] to predict the 3D mesh parameters of multiple people. Since the center position of each person is obtained, the body parameters of each person can be effectively extracted even in the case of mutual occlusion by multiple people. Moreover, combining body center points and body meshes is robust to complex situations compared to traditional learning methods.

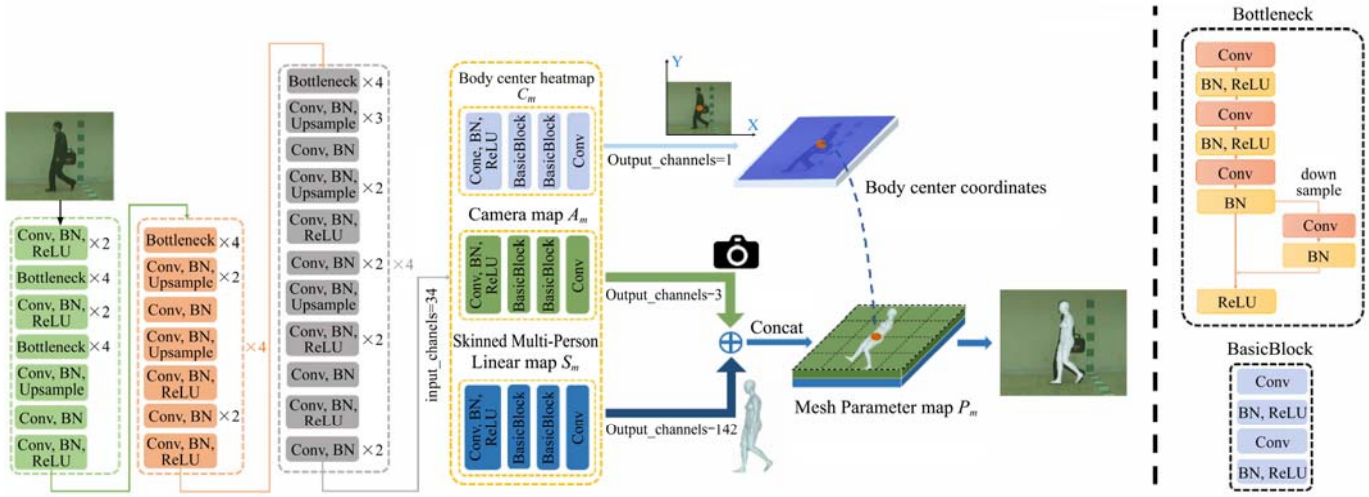

Figure 3 One-stage fashion for Multiple 3D People (ROMP) first takes the resnet network as backbone. Input an image containing a human body, the method uses three branch networks to predict three different maps, the first is the Body Center heatmap, which gets the 2D information of each person's body center position, and the second is the Camera map, which describes the relationship between displacement, body shape, distance and depth of a person, the third is a Skinned Multi-Person Linear map (SMPL), which describes the 3D pose and shape information of a human body, and the Skinned Multi-Person Linear map establishes pose and efficient mapping of shapes to human body 3D mesh information. Mesh Parameter map combines Camera map and Skinned Multi-Person Linear map, including 3D grid information of human body and displacement distance information. Finally, the final 3D mesh information of the human body is estimated by combining the body center heat map and the Mesh Parameter map

Usage of 3D Point Cloud Data. Even if there is no ground truth value, the parameterized body model SMPL obtained from the 3D human body pose shape estimation network can efficiently model the 3D information of the person, and can effectively extract the gait feature information. Different parameters in SMPL contain different information of the human body. By visualizing each parameter, we can observe the following features:

• cam (1, 3): Denotes the camera parameters of a weak perspective camera, represented by $\theta \in \mathbb{R}^3$.

• body_pose (1, 69): Represents the SMPL pose parameters, corresponding to $\theta \in \mathbb{R}^{69}$.

• smpl_betas (1, 10): Signifies the SMPL shape parameters, represented by $\theta \in \mathbb{R}^{10}$.

• joints (71, 3): Refers to the 3D pose results, represented by $\theta \in \mathbb{R}^{71\times3}$.

• verts (6890, 3): Represents the 3D coordinates of the human body mesh in 3D, corresponding to $\theta \in \mathbb{R}^{6890\times3}$.

• pj2d (71, 2): Denotes the 2D coordinates of the keypoints filled in the input image, represented by $\theta \in \mathbb{R}^{71\times2}$.

• pj2d_org (71, 2): Signifies the 2D coordinates of the keypoints in the original input image, corresponding to $\theta \in \mathbb{R}^{71\times2}$.

• can_trans (1, 3): Represents the rough 3D translation obtained from the estimated camera parameters conversion, represented by $\theta \in \mathbb{R}^3$.

• centers_conf (1, 1): Denotes the confidence value of detecting the person on the center image, corresponding to $\theta \in \mathbb{R}^1$.

Visualizing each of these parameters provides valuable insights into the different aspects of the system, such as camera settings, pose estimation, shape representation, keypoint localization, and confidence evaluation. These visualizations aid in understanding the underlying processes and can facilitate further

analysis and improvement of the overall system.

Due to the limited dimensionality of bone points, it may not have sufficient capacity to represent the complete features of the human body. As a solution, we utilize $\theta \in \mathbb{R}^{6890\times3}$ 3D human body mesh, which represents the 3D point clouds for data processing. The use of point cloud information provides a rich and extensive representation, which proves to be suitable for extracting gait recognition features. The larger dimensionality of the point cloud data allows for more comprehensive and detailed feature extraction, enhancing the accuracy and effectiveness of gait recognition algorithms.

The three-dimensional human body mesh information can be visualized using matplotlib.pyplot library, resulting in a scatter plot as shown in Figure 4. To normalize the data and make it usable, Min-Max Scaling technique can be applied:

$$X = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

where, X represent the coordinate value; $X_{\min}$ represent the minimum value of the coordinate, and $X_{\max}$ represent the maximum value of the coordinate.

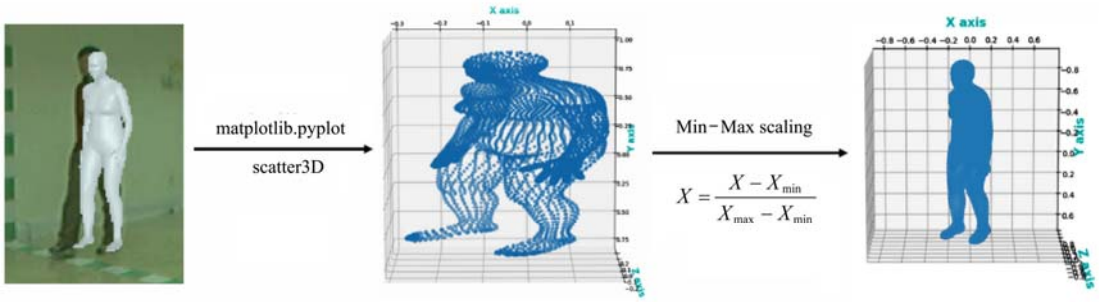

Figure 4 To visualize the three-dimensional human body mesh information using matplotlib.pyplot and subsequently normalize the data using Min-Max Scaling

This scaling process ensures that the data is normalized within a specific range. By performing Min-Max Scaling, the three-dimensional point cloud data can be effectively processed and utilized for further analysis or applications.

In this study, we extracted three-dimensional point cloud information from video data collected from a cohort of 124 individuals. Each individual was recorded in three different states, capturing variations in gait patterns. Additionally, the recording setup included 11 different camera angles, providing multiple perspectives for data collection.

By analyzing the video data and applying appropriate computer vision techniques, we were able to extract and reconstruct three-dimensional point clouds representing the spatial coordinates of the human body in each frame. This rich dataset allowed for a comprehensive analysis of gait patterns, considering variations across individuals, states, and camera angles.

3.2 Map 3D point cloud to 2D gait features

Since the reconstructed 3D point cloud contains a large amount of 3D points, bringing huge computational cost and redundant points, we need to map 3D point cloud to 2D feature spaces to reduce computational cost and preserve as many essential gait features as possible.

The mapping method includes 2D mapping by coordinate simplification and autoencoder mapping. We will introduce the two kinds of mapping methods in the following section.

2D mapping by coordinate simplification. The data representation of human body is 3D point cloud, which occupies a large amount of storage space and has a large computational cost, which is not convenient for data processing and calculation. If the 3D point cloud is projected to 2D form, it can save a lot of space storage and computing costs, and it is also very convenient for feature representation and matching. In this paper, a 2D form of 3D human point cloud is proposed. As shown in Figure 5, Without losing information, a more simplified form is used to represent information, which makes it possible to store and recognize massive 3D human bodies.

We propose a projection function T that maps the 3D coordinates of all subjects from 1 to n to 2D coordinates, resulting in a set of N_θ results. The p represent the 3D coordinate.

$$N_\theta = T(P^1_{(x,y,z)},...,P^n_{(x,y,z)})$$

For a given set of 3D point cloud data, one coordinate dimension is systematically removed, leaving the remaining two dimensions. Subsequently, the data is graphically represented, with a black background and white foreground, resulting in a binary image. This process is iteratively repeated by successively eliminating the other two coordinate dimensions, thereby yielding three distinct data representations.

In terms of methods, there have been studies using depth images to represent 3D human body structures. However, this method can only store the spatial structural information of the human body, while it can not represent the texture information that contains rich identification features, which makes depth images difficult to further improve the recognition performance. In fact, human information includes both structural features and texture features. Both structural and texture information can describe human physiological features. Comprehensive usage of these two features can better extract human identification information and improve the recognition performances. From the perspective of information collection equipment, the current 3D human collection equipment can basically carry color texture and acquiring 3D human structure information. Therefore, if the two modes of human 3D structure information and texture information can be used simultaneously, more complete human identification features can be extracted, and the overall performance of the human identification system can be improved through multi-modal identification fusion.

Autoencoder mapping. An autoencoder is an unsupervised learning neural network that encodes input data into a low-dimensional vector, and then decodes it into the original input through a decoder. Autoencoders are used in many fields, including dimensionality reduction of features. The traditional method of feature dimensionality reduction is principal component analysis (PCA), but the dimensionality reduction performance of autoencoders is better than PCA. Since the neural network can extract robust features more effectively, the decoder can

reconstruct the encoded vector as input, which can better obtain the encoded features, and the encoder can also be used for classification tasks, so it shows that the encoder It can achieve the function of effective feature extraction.

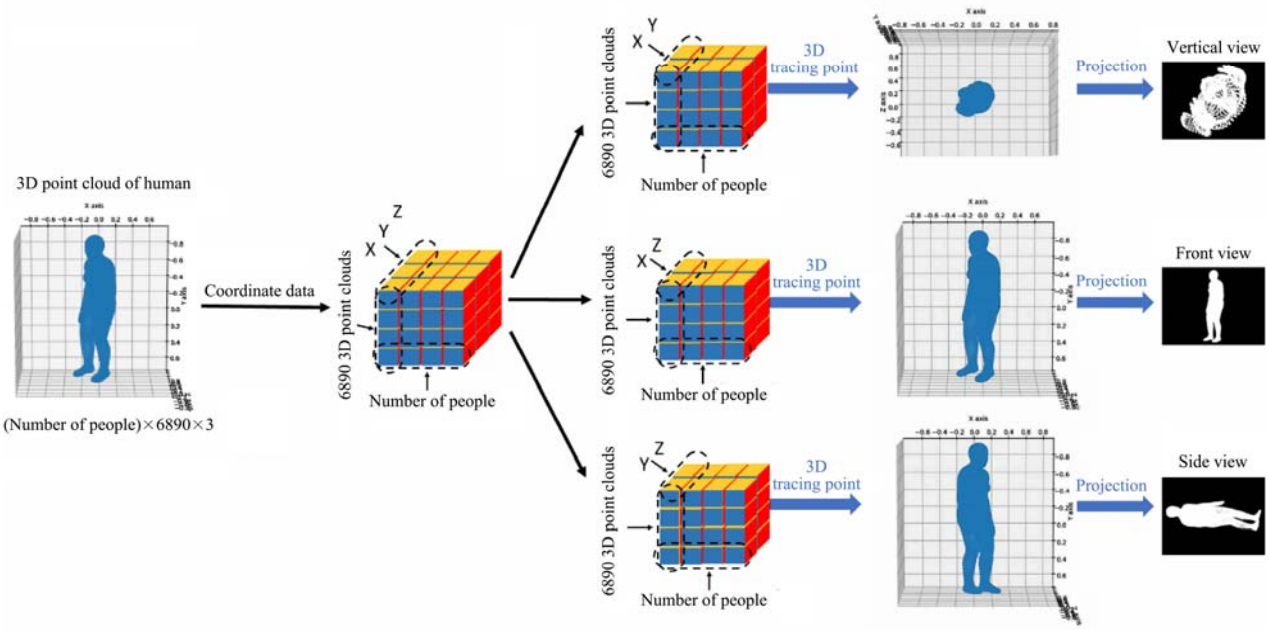

Figure 5 Mapping 3D point cloud to 2D binary images. We first take out the coordinates of a dimension to get the 2D cutting plane formed by the other two dimensions, then get the binary images

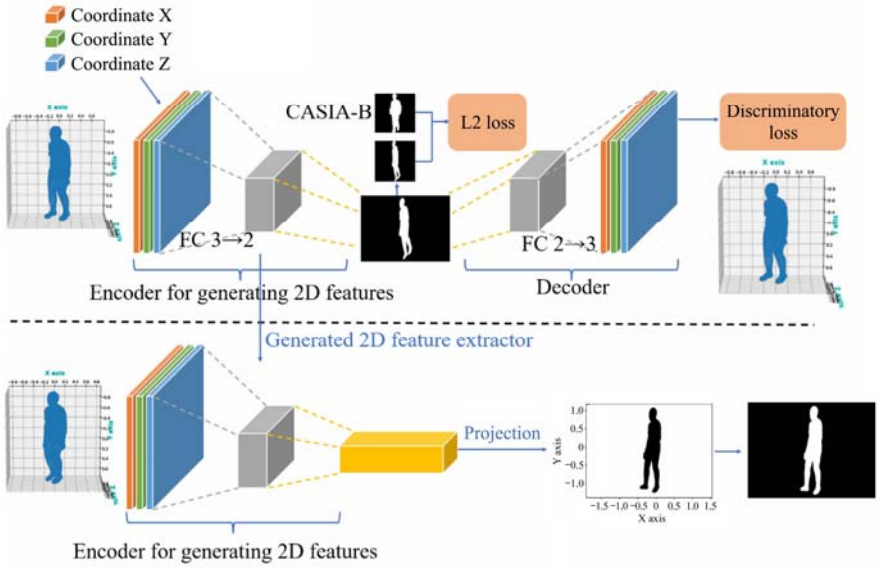

Figure 6 The 3D point cloud data is taken as input of the encoder E network in the data format of each 3D point. After encoder E reducing the dimension from 3D to 2D, the decoder G decodes the 2D data to 3D, where the 3D reconstruction loss is used to force the decoded 3D data to fit the original 3D data. Then we input the 3D point cloud data into the generated encoder E, and then use a discriminator D to determine whether the projected 2D data is consistent with the natural binary images, and then output the data that is judged to be true.

The Autoencoder consists of two main parts: the encoder is used to encode the input, and the decoder uses the encoded features to reconstruct the input, which can be defined as ϕ and ψ

$$\phi : \chi \to \gamma$$
$$\psi : \gamma \to \chi$$
$$\phi, \psi = \arg\min_{\phi,\psi} \| \chi - (\psi \circ \phi)\chi \|$$

Given a hidden layer, the input $x \in \mathbb{R}^d = \chi$ is accepted from the encoding stage of the encoder and mapped to $h \in \mathbb{R}^p = \gamma$:

$$h = \sigma(Wx + b)$$

where, h represents the feature vector; σ is the activation function; W is the weight matrix, and b is the offset vector. The weight and offset vectors are updated through backpropagation as the training progresses. Mapping h from the decoding stage of the encoder to the reconstructed x' (consistent with the shape of x):

$$x' = \sigma'(W'h + b')$$

The σ', W', b' of the decoder may be independent of the σ, W, b of the encoder.

Autoencoder is to minimize square error reconstruction loss:

$$\mathcal{L}(x, x') = \frac{1}{n}\sum_{i=1}^{n} \| x_i - x_i' \|^2$$
$$= \frac{1}{n}\sum_{i=1}^{n} \| x_i - \sigma'(W'(\sigma W x_i + b)) + b' \|^2$$

The model is trained using a composite loss function that

incorporates both L2 loss and 3D discriminative loss, with weighted hyperparameters α and β respectively:

$$\mathcal{L} = \alpha \mathcal{L}_{L2} + \beta \mathcal{L}_{dis}$$

Similar to the training mode of other neural networks, the training of Autoencoder also uses backpropagation to update parameters to converge the network. When the dimension of the intermediate feature γ is smaller than that of the input data χ, the intermediate feature $\phi(x)$ can be regarded as the feature compression of the input data x, which is used for dimensionality reduction of the feature. If the dimension of the intermediate feature is greater than or equal to the dimension of the input number, the autoencoder may learn the identity function or useless function representation, but studies have shown that the encoder can still learn useful feature information. Therefore, the dimension and number of layers in the autoencoder structure can be reasonably designed according to the form of the task.

The construction principle of a dual loss function typically involves the design of two main components, aiming to simultaneously optimize two distinct objectives during the training process. This type of loss function design is commonly employed in deep learning tasks, where one loss function is utilized to optimize the primary task, while the other is often employed to optimize an auxiliary task or provide regularization.

The primary task discriminative loss function is typically tailored to the model's primary objective. This loss function measures the model's performance on the primary task, guiding the updates of model parameters to enhance its capability in fulfilling the primary task. The auxiliary task L2 loss function is employed to optimize an auxiliary objective related to the primary task. This objective may be slightly different from the primary task or serve as a regularization term to improve the model's generalization performance.

The overall construction of the dual loss function usually involves a weighted sum of the primary task loss function and the auxiliary task loss function. This allows the model to simultaneously optimize both loss functions during training, enabling a more comprehensive learning of task-relevant features. This design is effective in multi-task learning scenarios or when leveraging auxiliary tasks to guide learning and improve overall model performance.

In our proposed Autoencoder mapping of 3D point cloud data, we consider that the generated 3D data is huge, and we need to map the 3D point cloud of each image to 2D data, so our encoder network structure needs to be simple enough, so we design an effective fully connected neural network as the structure of the encoder and decoder. The use of discrimination loss and L2 loss forces the self encoder structure to effectively extract two-dimensional features.

The incorporation of both discriminative loss and L2 loss in the model aims to effectively extract 2D features within the autoencoder architecture. The discriminative loss measures the discrepancy between the generated 3D outputs and the original input 3D data, compelling the network to focus on improving the 3D modeling performance. Meanwhile, the L2 loss evaluates the dissimilarity between the encoder-generated 2D data and the black-and-white contour images from CASIA-B dataset, promoting the convergence of the network towards desired 2D outcomes. By employing these distinct loss functions, the autoencoder structure is able to extract 2D features more efficiently.

The combined usage of discriminative loss and L2 loss facilitates the autoencoder structure in effectively capturing and reconstructing high-quality 2D features. This comprehensive approach of incorporating multiple loss functions enhances the performance of the autoencoder and yields improved results in 2D modeling tasks.

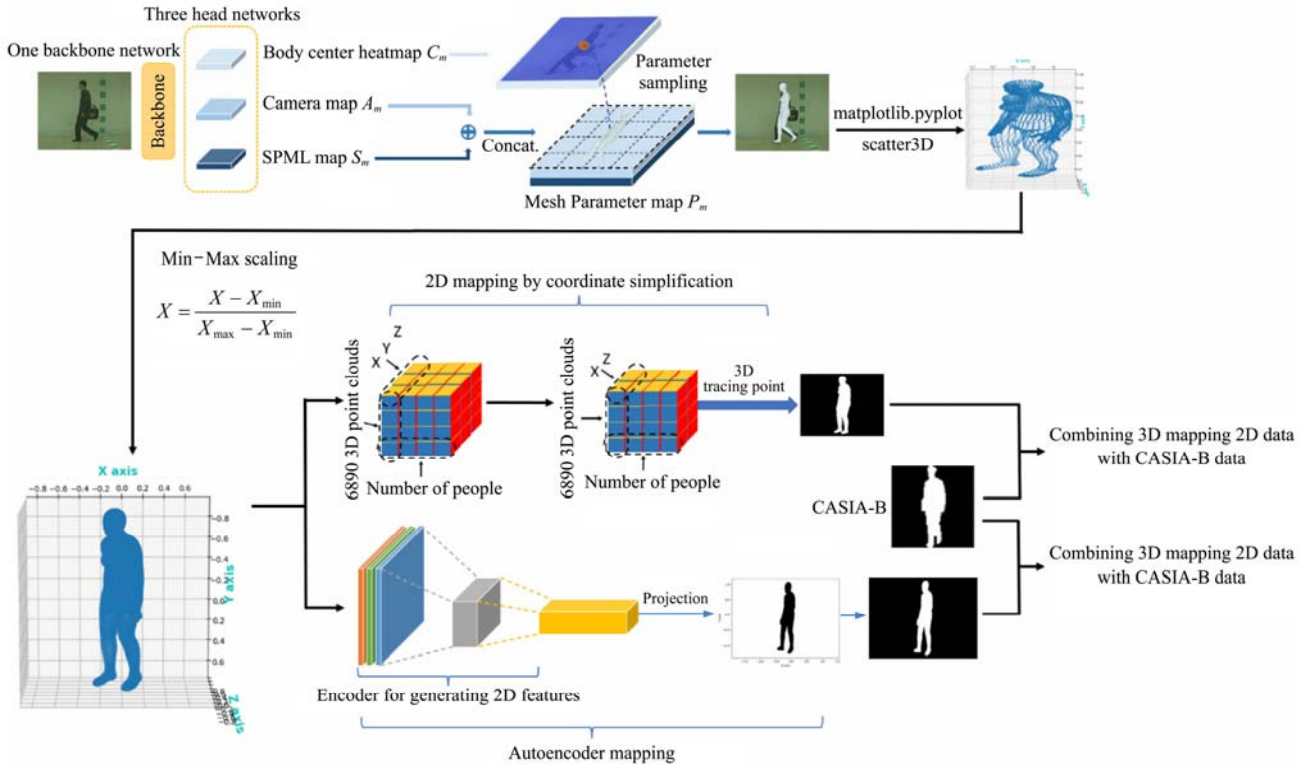

Figure 7 The process of mapping 3D point cloud data to 2D images can be visualized through a flowchart. The upper branch represents the direct 2D mapping approach applied to process the 3D data, while the lower branch represents the autoencoder mapping approach. Eventually, the resulting data is combined with the CASIA-B dataset, creating a combined database that incorporates both 3D and 2D information

Mapping 3D point cloud to 2D gait images. To arrange the 3D point cloud data of different subjects under different walking conditions, we generate 3D point cloud from different perspectives of different states of each person, get 2D point data from 3D point cloud data through 2D mapping, normalize the 2D data, retain the morphological features of the human body, represent the 2D data in form of images, and convert the images to 64×64 binary contour images by clipping, then merge them with gait samples in the original dataset.

By applying either the 2D mapping or the autoencoder mapping to process the binary black-and-white contour images, we can align the same states and angles of the participant with the corresponding states and angles in the CASIA-B dataset. This integration involves combining the processed data from both approaches, ensuring that the same conditions and perspectives are accurately represented. This combined dataset facilitates comparative analysis and evaluation between the participant's data and the reference data from CASIA-B, enabling a comprehensive examination of similarities and differences in various states and angles.

3.3 Gait recognition based on gait features obtained from 3D point cloud

The 2D gait images obtained from 3D point cloud have several advantages: (1) the 3D point cloud are built based on essential human structures which are much less affected by dress and carrying, then the projected 2D gait images are more robust to dress and carrying variances; (2) the 3D point cloud can be projected to any view planes to obtain 2D gait images under arbitrary views, which is helpful in expanding adequate gait samples for training gait classification networks; (3) the 3D point cloud contains a large amount of dense features points describing more comprehensive human moving characteristics than sparse key joint points.

After projecting 3D point cloud to 2D gait images, the 2D gait images are input into We adopt the method of treating gait as a sequence, which can effectively preserve the connection between human walking gaits. We put the combined multiple frames which are processed from 3D point cloud data into the gait method. The image consists of a continuous sequence of 30 frames capturing the gait over time, which is not affected by the order of the input frames. The gait sequence can be effectively integrated for videos taken under different conditions from different perspectives, clothing and carrying items. It has a more precise and powerful effect in practical application scenarios.

Apart from GaitSet method, there have also appeared other methods considering multiple gait frames as a set, like GaitPart and GaitGL. In this paper, we will test the effectiveness of our gait samples projected from 3D point cloud on different set-based classification networks.

GaitPart used the Focal Convolution Layer to extract enhanced gait features, and also uses multiple parallel Micro-motion Capture Module (MCM) to focus on the short-term feature learning of the gait, thereby improving the overall performance of gait recognition.

GaitGL proposed to combine global and local gait information. This module is composed of multiple global and local modules to effectively extract feature information for gait recognition. In addition, it also uses A Local Temporal Aggregation (LTA) that preserves spatial information by reducing temporal information to learn more spatially.

4 Experiments

In this section, We present the experimental comparison results between our method and GaitSet[13], GaitPart[15] and GaitGL[65]. The two 2D feature extraction methods: direct two-dimensional mapping and self encoder mapping are compared. To illustrate the effectiveness of each experimental setting option, we provide a series of detailed ablation studies and standard measurements.

4.1 Datasets and Training Details

CASIA-B. The CASIA-B dataset[66] is the most widely used gait recognition dataset in current research. This data set contains 124 different people. Each person has 3 walking states, which are normal, carrying a bag and wearing a coat. Among them, there are 6 groups of videos in the normal state, 2 groups of videos in the carrying bag, and 2 groups of videos in the coat. Each person has a total of 10 groups of videos, each group of videos contains an average of 11 camera angles from 0 to 180 degrees, and the interval sampling angle is 18 degrees, so the there are 13640 gait vidoes in the dataset. The data set can be divided into different training sets and test sets correspond to different tasks.

Figure 8 From the top left to the bottom right are clips from different angles of the subject's video in the CASIA-B[66] gait dataset

As shown in Table 1, the first 74 people are selected as the training set to train the network model, and the remaining 50 people are taken as the test set [41]. In the test, we set the gallery to NM\#01-04 and the probe to NM\#5-6, BG\#1-2, CL\#1-2 to test the experimental effects of three different states: normal (NM), carrying a bag (BG) and wearing a coat (CL).

Table 1 Experimental settings of CASIA B dataset

Training	Test	
	Gallery set	Probe set
ID: 001-074 Seqs: NM01-NM06 BG01-BG02, CL01-CL02	ID: 075-124 Seqs: NM01-NM04	ID: 075-124 Seqs: NM05-NM06 BG01-BG02, CL01-CL02

Training Details. In all experiments, the input is a set of binary human silhouette images with image size 64×44. This data is directly combined by the CASIA-B dataset and our proposed 3D point cloud dataset. The optimizer is Adam, the learning rate is set as 1e-4, the weight decay is set as 5e-4. The batch size is set as (8, 16). Also, our trained model is 150K iterations. In the training process, the gait sequence frame number of the input data is 30. In the testing phase, the entire gait sequence frames of a person are input into the deep model. The model is trained and tested using 4 NVIDIA 3080 GPUs.

4.2 2D mapping VS Automatic encoder mapping

The generated results of the two different mapping approaches for the 2D binary images are depicted in the accompanying Figure 9. It is evident that the 2D mapping technique successfully preserves the original view mapping effect, resulting in generated images that closely resemble the visual characteristics of the original data. On the other hand, the autoencoder-based mapping method demonstrates the ability to capture the features of the underlying 3D information, effectively aggregating global features and exhibiting discernible distinctions in local features. This signifies the proficiency of the autoencoder in learning multi-level representations of the input data. The provided figure visually illustrates the disparities between the two mapping approaches and their corresponding effects.

Figure 9 Visualization of different 2D mapping methods: The top row displays CASIA-B data, the middle row represents direct 2D mapping, and the bottom row represents autoencoder mapping. These columns correspond to different perspectives. From the visualizations, it is evident that the autoencoder method excels in aggregating global features and extracting local features

Both 2D mapping and automatic encoder mapping from 3D point cloud have advantages and limitations.

First, we conduct 2D mapping experiments on GaitSet, GaitPart and GaitGL. Compared with many features based on appearance and skeletal joint points, the 3D point cloud features extracted by the proposed method are large and effective, and the occlusion information for human effect is less. Our method shows that the adequate gait features existing in 3D point cloud are capable of bringing improvement in different walking conditions. The proposed method uses the binary contour maps of 3D point cloud, which proves the superiority of applying 3D point cloud in gait recognition tasks. The experimental results show that the embedding of 3D point cloud information can effectively improve gait recognition performances.

We have compared our proposed method with the most advanced gait recognition methods in terms of clothing texture confusion and mixing in Table 2, We can observe that, inputting the two-dimensional mapping data in the GaitSet method can achieve superior experimental results compared with the original method in terms of clothing texture confusion. Meanwhile, our data has achieved competitive performances on GaitPart (2Dmapping) nm \# 5-6.

Under these conditions, the accuracy of our proposed method based on 3D point cloud is satisfactory, and the accuracy under NM condition is higher than that of the original method. The experimental results show that the proposed 3D point cloud method has more obvious advantages in the effect of gait recognition, and can propose powerful features that are more suitable for classification, and has a wide range of application scenarios in the field of recognition.

We also carry out experiments based on the Autoencoder mapping method of 3D point cloud. It can be found in Table 3 that, compared with the 2D mapping method, the accuracy of the Autoencoder method is higher. The reason is that, autoencoder mapping can map 3D point cloud to 2D features by neural networks, this process can preserve more essential gait features than 2D mapping based on dimension reduction.

In gait classification models, the average performances based on autoencoder mapping are improved. Experiment results demonstrate that the features extracted from the autoencoder can be better applied to gait recognition task.

We have compared the proposed model with the most advanced gait recognition methods. From Table 3, the average results of our data in all cases of GaitSet (Autoencoder) are obviously better than the experimental results in the original method. Meanwhile, our model has achieved competitive performance on BG \# 5-6 / CL \# 1-2. The experimental results are higher than those in the original methods, which indicates that the method of 3D point cloud can be applied to the effective extraction of gait features.

We also test the experimental results of 3D point cloud mapping by self encoder on GaitPart. The average accuracy of the experimental results under NM state is improved. The experimental results show that the method has obvious advantages under NM conditions, indicating that the model can extract more robust gait features under normal walking.

On the GaitGL method, the average accuracy of our experiment has improvement. The experimental results show that the method has obvious advantages under NM conditions.

It can be seen from Table 2 and Table 3 that, under these conditions, the average recognition accuracy of the proposed method has been improved under NM conditions, indicating that the method has advantages under NM conditions, and can extract more robust gait features, leading to improved experimental results.

At the same time, when combining Autoencoder mapping with the gait recognition method, the recognition accuracies under each state, including NM, BG and CL, have been improved, which comprehensively exceeds the original method, indicating that 3D point cloud data can significantly improve the robustness of extracted features, effectively reduce the impact of clothing texture features on the recognition accuracy, and thus have a good recognition effect.

Table 2 Experimental results of various models based on 2D mapping by coordinate simplification.

Gallery NM#1-4		0°-180°											mean
	Probe	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitSet(2Dmapping)	89.9	97.9	99.2	96.3	91.8	90.6	94.6	**98.3**	97.9	96.1	**87.2**	94.5
	GaitPart	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitPart(2Dmapping)	94.0	98.6	**99.4**	**98.6**	**95.4**	92.2	95.5	**98.7**	99.2	**98.0**	**90.7**	**96.4**
	GaitGL	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitGL(2Dmapping)	94.2	97.0	99.0	97.3	95.7	93.1	96.8	98.5	98.9	97.4	90.2	96.2
BG #1-2	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitSet(2Dmapping)	**85.7**	**92.3**	**93.4**	**90.2**	**85.2**	80.6	**86.1**	**90.7**	**94.1**	92.8	**80.8**	**88.4**
	GaitPart	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitPart(2Dmapping)	87.5	92.9	94.3	92.2	86.9	82.9	87.4	93.4	93.4	90.8	82.4	89.5
	GaitGL	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitGL(2Dmapping)	88.4	93.6	95.2	93.4	91.9	88.1	92.0	95.9	96.5	93.7	87.2	92.4
CL #1-2	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitSet(2Dmapping)	**67.7**	**79.5**	79.9	**75.5**	68.5	67.1	70.8	72.5	**75.3**	**73.8**	**59.4**	**71.8**
	GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitPart(2Dmapping)	67.0	78.0	82.4	79.1	73.3	68.6	73.6	79.6	79.0	76.3	63.2	74.6
	GaitGL	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitGL(2Dmapping)	70.4	83.8	87.9	85.9	81.4	74.6	80.0	82.0	82.1	79.7	65.0	79.3

Table 3 Experimental results of various models mapped by autoencoder.

Gallery NM#1-4		0°-180°											mean
	Probe	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitSet(Autoencoder)	**92.1**	**98.8**	99.1	**97.2**	**94.4**	**92.9**	**96.5**	**97.9**	98.6	**97.8**	**90.1**	**95.9**
	GaitPart	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitPart(Autoencoder)	93.2	98.3	99.0	98.3	**94.7**	**92.8**	95.9	98.2	99.1	**98.5**	**92.6**	**96.4**
	GaitGL	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitGL(Autoencoder)	95.5	**99.0**	**99.2**	**98.0**	96.1	94.6	**97.6**	98.9	99.1	**99.1**	**94.5**	**97.5**
BG #1-2	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitSet(Autoencoder)	**87.6**	**93.3**	**94.5**	**90.7**	**88.3**	**81.2**	**84.7**	**92.4**	**95.9**	93.1	**83.2**	**89.5**
	GaitPart	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitPart(Autoencoder)	**89.9**	94.2	95.1	92.7	**88.6**	83.5	88.2	93.3	94.2	**93.5**	**86.3**	90.9
	GaitGL	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitGL(Autoencoder)	92.4	96.3	96.7	94.0	93.1	88.7	91.3	95.9	97.4	95.2	90.5	93.8
CL #1-2	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitSet(Autoencoder)	**68.5**	**80.6**	80.1	**77.8**	70.6	68.5	70.3	**75.0**	**76.9**	**75.2**	**58.3**	**72.9**
	GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitPart(Autoencoder)	70.7	81.7	85.5	**83.8**	75.9	**72.9**	**77.9**	81.0	81.8	**81.1**	**67.2**	78.1
	GaitGL	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitGL(Autoencoder)	75.0	87.9	89.5	86.1	84.5	**79.8**	82.0	84.6	84.4	81.9	67.8	82.1

4.3 Comparison with State-of-art Methods

In this section, we compare the proposed method with eight advanced technologies with the same experimental settings, namely CNN-LB, skeletonGait, GEINet, deepCNNs, poseGait, SDHF-GCN, GaitPart and GaitSet. The average recognition accuracies under the walking condition NM, BG and CL are shown in Table 4.

To ensure a systematic and comprehensive comparison with other existing research methods, we evaluate all the experimental results of NM, BG and CL. Except for CNN-LB[41], which is based on the gait energy map, the others are gait recognition methods based on gait sequences. It can be seen that most of the methods with better results are based on gait sequences, so it also shows that gait sequences can Better express the robust characteristics of human walking gait. In the embedding of our 3D point cloud data, in the case of GaitGL's normal walking NM, it has reached the highest accuracy. In the case of wearing a coat CL, it also improves at individual angles, so our method effectively improves, the effect of gait recognition also proves that 3D point cloud data can have great application potential in gait recognition field.

Experimental findings demonstrate a significant improvement in angle precision for certain pedestrian states. This improvement is attributed to the ability of mapped 3D data to effectively eliminate occlusions, enabling the accurate extraction of human body features. However, suboptimal results are observed for specific angles, primarily due to inherent inaccuracies in the 3D

modeling algorithm, resulting in insufficient precision in human feature extraction.

We can see that the recognition rate of the proposed method is higher. Especially, this method is capable of obtaining high recognition rates in case of clothing changes. This means that our proposed method has a better effect on removing occlusions, which is also the advantage of 3D point cloud features. The 3D point cloud data can model essential human body structures regardless of clothing changing, while the appearance-based features often change due to the clothing. Then we can see that 3D point cloud can inherit the advantages of both model-based and appearance-based methods.

Table 4 Comparison of identification accuracy (%) with the most advanced methods.

Gallery NM#1-4		0°-180°											mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	CNN-LB	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	SkeletonGait	82.3	87.5	91.7	91.3	89.0	88.5	89.1	90.5	92.7	91.0	82.4	88.7
	GEINet	40.2	38.9	42.9	45.6	51.2	42.0	53.5	57.6	57.8	51.8	47.7	48.1
	DeepCNNs	77.3	82.8	85.1	86.0	85.5	85.4	83.7	81.5	80.5	83.9	77.6	82.7
	PoseGait	55.3	69.6	73.9	75.0	68.0	68.2	71.1	72.9	76.1	70.4	55.4	68.7
	SDHF-GCN	77.3	82.8	85.1	86.0	85.5	85.4	83.7	81.5	80.5	83.9	77.6	82.7
	GaitGraph	85.3	88.5	91.0	92.5	87.2	86.5	88.4	89.2	87.9	85.9	81.9	87.7
	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GaitGL	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	99.3	98.8	94.0	97.4
	GaitSet(Autoencoder)	92.1	98.8	99.1	97.2	94.4	92.9	96.5	97.9	98.6	97.8	90.1	95.9
	GaitPart(Autoencoder)	93.2	98.3	99.0	98.3	94.7	92.8	95.9	98.2	99.1	98.5	92.6	96.4
	GaitGL(Autoencoder)	95.5	**99.0**	99.2	98.0	96.1	94.6	**97.6**	**98.9**	99.1	**99.1**	**94.5**	**97.5**
BG #1-2	CNN-LB	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	SkeletonGait	73.8	80.6	80.7	81.2	77.3	75.5	78.5	76.3	80.8	79.0	74.2	78.0
	GEINet	34.2	29.3	31.2	35.2	35.2	27.6	35.9	43.5	45.0	39.0	36.8	35.72
	DeepCNNs	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	PoseGait	35.3	47.2	52.4	46.9	45.5	43.9	46.1	48.1	49.4	43.6	31.1	44.5
	SDHF-GCN	67.5	73.9	73.2	74.3	68.5	68.5	70.5	69.0	62.2	68.7	60.1	68.8
	GaitGraph	75.8	76.7	75.9	76.1	71.4	73.9	78.0	74.7	75.4	75.4	69.2	74.8
	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart	89.1	94.8	96.7	95.1	88.3	94.9	89.0	93.5	96.1	93.8	85.8	91.5
	GaitGL	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	GaitSet(Autoencoder)	87.6	93.3	94.5	90.7	88.3	81.2	84.7	92.4	95.9	93.1	83.2	89.5
	GaitPart(Autoencoder)	89.9	94.2	95.1	92.7	88.6	83.5	88.2	93.3	94.2	93.5	86.3	90.9
	GaitGL(Autoencoder)	92.4	96.3	96.7	94.0	93.1	88.7	91.3	95.9	97.4	95.2	90.5	93.8
CL #1-2	CNN-LB	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	SkeletonGait	60.7	62.6	68.1	68.5	65.5	65.9	64.8	65.9	67.4	64.6	58.9	64.8
	GEINet	19.9	20.3	22.5	23.5	26.7	21.3	27.4	28.2	24.2	22.5	21.6	23.5
	DeepCNNs	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	PoseGait	24.3	29.7	41.3	38.8	38.2	38.5	41.6	44.9	42.2	33.4	22.5	36.0
	SDHF-GCN	63.4	65.4	66.7	64.8	63.0	66.2	69.1	63.3	61.1	65.9	60.7	64.5
	GaitGraph	69.6	66.1	68.8	67.2	64.5	62.0	69.5	65.6	65.7	66.1	64.3	66.3
	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GaitGL	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	GaitSet(Autoencoder)	68.5	80.6	80.1	77.8	70.6	68.5	70.3	75.0	76.9	75.2	58.3	72.9
	GaitPart(Autoencoder)	70.7	81.7	85.5	83.8	75.9	72.9	77.9	81.0	81.8	81.1	67.2	78.1
	GaitGL(Autoencoder)	75.0	87.9	89.5	86.1	**84.5**	**79.8**	82.0	84.6	84.4	81.9	67.8	82.1

4.4 Ablation Experiments

Influence of 3D point cloud reconstruction method. When using different 3D human pose shape estimation methods to reconstruct 3D point cloud data, the modeled human body structures are also different. We have done experiments of another human_dynamic method[67] on the 3D human body pose estimation. This three-dimensional modeling is an earlier research.

The generated 3D point cloud is converted into 2D binary contour maps. The obtained 2D binary contour maps are merged with original gait images in CASIA-B to expand gait samples. The experimental results in Table 5 are obtained.

For better 3D point cloud reconstruction effect, one advanced method ROMP is adopted. The 3D point cloud data generated by ROMP is converted into 2D binary contour maps, and then training and testing process are conducted on GaitSet, GaitPart and GaitGL. It can be found from the results in Table 6 that, using one of the most advanced 3D models can improve the final accuracy. Moreover, we can find that the recognition accuracies in Table 6 are much lower than the optimized recognition accuracies in Table 7. The reason is, the results in Table 6 have taken all 3D point cloud data as input, since more 3D point cloud data will generate more different distribution characteristics with the original gait samples in existing datasets, which becomes harder for deep networks to fit more complexed gait samples mapped from 3D point cloud data.

Table 5 Comparison of identification accuracy (%) with the most advanced methods.

Gallery NM#1-4		0°-180°											mean
	Probe	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet(human_dynamic)	90.8	96.9	99.0	96.6	92.9	91.6	94.2	96.9	97.6	96.1	87.5	94.6
	GaitPart(human_dynamic)	83.9	90.9	94.6	94.0	89.0	87.3	90.2	95.2	95.2	91.9	78.1	90.0
	GaitGL(human_dynamic)	80.3	90.5	93.0	92.7	91.6	89.7	90.8	95.3	94.2	91.9	74.5	89.5
BG #1-2	GaitSet(human_dynamic)	86.2	91.3	93.6	90.1	84.4	79.5	84.7	91.0	94.2	91.0	81.1	87.9
	GaitPart(human_dynamic)	72.4	79.9	83.9	79.6	78.7	73.2	76.9	81.5	83.8	75.4	62.1	77.0
	GaitGL(human_dynamic)	71.5	82.6	87.6	86.8	83.6	78.8	82.9	90.2	89.1	83.1	60.7	81.5
CL #1-2	GaitSet(human_dynamic)	57.2	72.0	77.6	76.2	69.6	67.5	69.3	71.5	72.1	64.9	49.1	67.9
	GaitPart(human_dynamic)	51.1	60.8	62.6	62.0	61.2	58.3	63.7	65.9	60.5	57.0	41.5	58.6
	GaitGL(human_dynamic)	43.8	63.0	71.5	73.4	69.3	62.9	70.6	74.0	68.9	57.9	34.0	62.7

Table 6 The gait recognition accuracy whose 3D point cloud is generated based on three-dimensional human pose estimation method called human_dynamic

Gallery NM#1-4		0° - 180°											mean
	Probe	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM #5-6	GaitSet(ROMP)	92.1	98.0	98.4	95.7	94.7	92.8	95.9	97.9	98.2	96.0	87.3	95.2
	GaitPart(ROMP)	83.4	91.7	94.7	91.8	88.8	88.0	89.5	92.6	93.2	91.0	75.8	89.1
	GaitGL(ROMP)	83.4	92.6	95.1	91.6	90.5	88.3	90.8	95.0	95.2	91.9	76.8	90.1
BG #1-2	GaitSet(ROMP)	86.7	93.0	94.1	89.3	84.5	79.9	84.1	90.6	94.9	90.2	81.3	88.0
	GaitPart(ROMP)	69.0	76.3	81.4	78.0	76.2	72.7	77.1	83.7	84.2	75.9	63.6	76.2
	GaitGL(ROMP)	69.6	83.7	90.1	90.0	84.8	82.6	86.9	91.3	92.2	83.1	63.9	83.5
CL #1-2	GaitSet(ROMP)	61.1	71.1	75.5	73.2	67.7	64.8	68.9	73.0	76.2	70.1	53.2	68.6
	GaitPart(ROMP)	47.4	63.5	66.1	66.8	62.5	60.1	67.3	68.5	66.7	60.1	48.8	61.6
	GaitGL(ROMP)	42.1	66.9	78.8	77.9	73.6	68.2	74.2	77.7	74.1	66.1	45.8	67.8

Influence of inaccurate 3D modeling frames. In 3D recognition, when the characters appear in the video and disappear in the video, inaccurate 3D modeling will occur in 3D recognition. In order to eliminate the negative impact of inaccurate modeling in all modeling data at the beginning, we remove the modeling samples of the first 10 frames and the last 10 frames, and train and test these gait samples on the original GaitSet, named GaitSet (01). The data accuracy is shown in Figure 10.

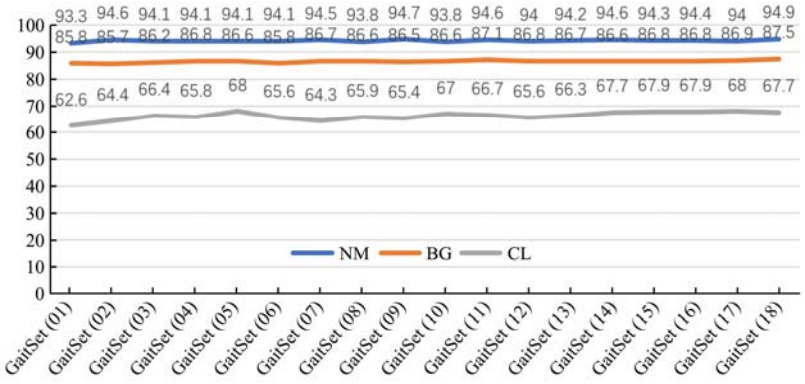

Figure 10 Accuracy based on the GaitSet method in different datasets. The amount of newly added 3D point cloud data has little impact on the NM and BG conditions, but has a certain range of fluctuations on the CL condition, and tends to be stable as a whole

In order to explore the impact of 3D data volume, we gradually reduce the generated 3D data in small batches and conduct training and testing on the original GaitSet. The definition of each data is given as follows: GaitSet (02) is 15 frames before and after the removal, and GaitSet (03) is 20 frames before and after the removal. After that, in order to explore the impact of data volume on accuracy in more details, and in the CASIA-B data set, it can be found that the characters are from right to left, and the picture occupation in the video increases with the video playing. Therefore, GaitSet (04) is set to remove the first 25 frames and the last 20 frames. Keep the number of frames removed from the back unchanged, and continue to remove the previous frames. GaitSet (05) removes the first 30 frames, GaitSet (06) removes the first 35 frames, GaitSet (07) removes the first 40 frames, GaitSet (08) removes the first 45 frames, GaitSet (09) removes the first 50 frames, GaitSet (10) removes the first 55 frames, GaitSet (11) removes the first 60 frames, GaitSet (12) removes the first 65 frames, and GaitSet (13) removes the first 70 frames, Gaitset (14) removes the first 75 frames, GaitSet (15) removes the first 80 frames, GaitSet (16) removes the first 85 frames, GaitSet (17) removes the first 90 frames, and GaitSet (18) removes the first 95 frames. The resulting accuracy data are shown in Figure 10. It can be found that the accuracy fluctuates. The amount of data from GaitSet (01) to GaitSet (14) decreases sequentially.

We have also done extensive experiments taking the GaitSet method as backbone, combining different amount of gait images obtained from 3D point cloud with original 2D binary contour maps of CASIA-B dataset. The data amount of projected gait images is larger than the original CASIA-B dataset. The accuracy can be improved to a certain extent by continuously reducing the

size of the projected gait images. This is due to the model's underfitting. When the data amount reaches a certain threshold, the accuracy can be improved.

Validity of 3D point cloud data. We combine the gait images projected from 3D point cloud data with the binary contour maps of the CASIA-B dataset. We can find that the accuracy of the NM based on the GaitPart method has been increased, which is higher than the original method. Based on the GaitSet and GaitGL method, there are also improvements in BG and Cl, which also proves the universality of the data. 3D point cloud data can eliminate the influence of carrying bags and wearing coats, and complete data modeling of the shape of the entire human body. The method of combining 3D point cloud and 2D binary data can achieve higher recognition accuracy.

4.4 Computational cost analysis

In the proposed method, most of the computational cost comes from the pre-processing process. That is to estimate the 3D human body point cloud from gait image sequences. We run the proposed method on a server equipped with 4 NVIDIA 3080 GPUs. and listed the time consumed in different steps in Table 7. For the three estimation steps, it takes about 0.1 seconds to convert the image to 3D point cloud data, 0.57 seconds for 2D mapping of 3D point cloud, and 0.0017 seconds for final clipping to 64×64. The calculation cost of 2D to 3D is actually low enough. It is obvious that the proposed method is fast and effective.

Table 7 The computational cost of different steps in the proposed method

Step	Time/s	Description
3D modeling	0.1	CPU & GPU
2D Mapping of 3D Point Cloud	0.57	CPU & GPU
2D feature clipping 64x64	0.0017	CPU

Yang Q G, Chen X, Lan Y B, Deng X L**.** Gait recognition based on 3D point cloud data augmentation This paper proposes a gait recognition method based on 3D point cloud, and demonstrate its effectiveness when the clothing and carrying textures becomes chaotic. The 3D point cloud data is mapped to two-dimensional feature representations insensitive to texture variances by autoencoder mapping. Specifically, we first obtain the 3D point cloud data reflecting essential human body structures, then map the 3D point cloud to 2D binary gait images of different views, then combine the mapped binary gait images with original gait samples to expand existing gait datasets. The experimental results suggest that the proposed method is effective in gait recognition tasks. This paper demonstrates that 3D point cloud is helpful in reducing the influence of dress and carrying and preserving adequate essential gait features, which is good at combining advantages of both appearance-based and model-based methods.

While the proposed method for gait recognition based on 3D point clouds demonstrates effectiveness, several limitations should be acknowledged. Firstly, challenges exist in acquiring and accurately annotating large-scale 3D point cloud datasets, potentially impacting the model's generalization and robustness. Additionally, mapping 3D point clouds to a two-dimensional feature representation using an autoencoder may result in information loss, affecting the preservation of fine details crucial for accurate gait recognition. The approach's performance in handling certain types or shapes of carried items could be limited, and the method’s adaptability to varying environmental conditions, such as lighting changes and different backgrounds, needs further validation. Furthermore, the extension of the dataset through the combination of mapped binary gait images with original gait samples requires comprehensive verification for its effectiveness and generalization. In conclusion, while the method shows promise in 3D gait recognition, thorough research and validation are necessary to address these limitations and ensure its applicability in diverse and complex scenarios.

6 Acknowledgment

This work was supported by Laboratory of Lingnan Modern Agriculture Project (Grant No. NT2021009), National Natural Science Foundation of China (Grant Nos. 61906074, 61675003, 61972178), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011276), the 111 Project (D18019), Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B020214003), Key-Area Research and Development Program of Guangzhou (Grant No. 202103000090), Key-Areas of Artificial Intelligence in General Colleges and Universities of Guangdong Province (Grant No. 2019KZDZX1012), the Fundamental Research Funds for the Central Universities (21620432), Key-Area Research and Development Program of Guangdong Province (2019B1515120010).

[References]

[1] L. Xue, M. Gao, C. Xing, R. Mart´ın-Mart´ın, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1179–1189.

[2] H. Zhu, Z. Zheng, and R. Nevatia. Gait recognition using 3-d human body shape inference, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 909–918.

[3] A. Xiao, J. Huang, W. Xuan, R. Ren, K. Liu, D. Guan, A. El Saddik, S. Lu, and E. P. Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9382–9392.

[4] M. Deng, Z. Fan, P. Lin, and X. Feng. Human gait recognition based on frontal-view sequences using gait dynamics and deep learning. IEEE Transactions on Multimedia, 2023.

[5] T. Teepe, A. R. Khan, J. Gilg, F. Herzog, S. H¨ormann, and G. Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. IEEE International Conference on Image Processing, 2021. 1, II-C

[6] G. Li, L. Guo, R. Zhang, J. Qian, and S. Gao. Transgait: Multimodal-based gait recognition with set transformer. Applied Intelligence, 2023, 53(2): 1535–1547.

[7] T. Huang, X. Ben, C. Gong, B. Zhang, R. Yan, and Q. Wu. Enhanced spatial-temporal salience for cross-view gait recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 1(1): pp.1–14.

[8] R. Liao, S. Yu, W. An, and Y. Huang. A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition, 2020. I, II-C, II-C, II-C.

[9] X. Liu, Z. You, Y. He, S. Bi, and J. Wang. Symmetry-driven hyperfeature gcn for skeleton-based gait recognition. Pattern Recognition, 2022, 125, pp. 1–13.

[10] F. Han, X. Li, J. Zhao, and F. Shen. A unified perspective of classification-based loss and distance-based loss for cross-view gaitrecognition. Pattern Recognition, 2022, 125, pp. 1–10.

[11] H. Qin, Z. Chen, Q. Guo, Q. M. J. Wu, and M. Lu. Rpnet: Gait recognition with relationships between each body-parts. IEEE Transactions on Circuits and Systems for Video Technology, 2022.

[12] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez. Exploring transformers for behavioural biometrics: A case study in gait recognition. Pattern Recognition, 2023, 143, p. 109798.

[13] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. National Conference on Artificial

Intelligence, 2018. I, II-B, IV.
[14] Y. Zhang, Y. Huang, S. Yu, and L. Wang. Cross-view gait recognition by discriminative feature learning. IEEE Transactions on Image Processing, 2020.
[15] 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. Computer Vision and Pattern Recognition, 2020. I, II-B, II-B, IV.
[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, 2023, pp. 1–22.
[17] V. Narayan, S. Awasthi, N. Fatima, M. Faiz, and S. Srivastava. Deep learning approaches for human gait recognition: A review, in 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE, 2023, pp. 763–768.
[18] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian. Multi-view gait image generation for cross-view gait recognition. IEEE Transactions on Image Processing, 2021.
[19] X. Chen, J. Weng, W. Luo, W. Lu, H. Wu, J. Xu, and Q. Tian,Sample balancing for deep learning-based visual recognition. IEEE Transactions on Neural Networks, 2020.
[20] L. Yao, W. Kusakunniran, Q. Wu, J. Xu, and J. Zhang. Collaborative feature learning for gait recognition under cloth changes. IEEE Transactions on Circuits and Systems for Video Technology, 2021.
[21] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu. Lidargait: Benchmarking 3d gait recognition with point clouds, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1054–1063.
[22] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and L. Cui. Gaitamr: Cross-view gait recognition via aggregated multi-feature representation. Information Sciences, 2023, 636, p. 118920.
[23] Y. Shi, L. Du, X. Chen, X. Liao, Z. Yu, Z. Li, C. Wang, and S. Xue. Robust gait recognition based on deep cnns with camera and radar sensor fusion. IEEE Internet of Things Journal, 2023.
[24] T. Teepe, J. Gilg, F. Herzog, S. H¨ormann, 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.
[25] 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.
[26] Z. He, W. Wang, J. Dong, and T. Tan. Temporal sparse adversarial attack on sequence-based gait recognition. Pattern Recognition, 2023, 133, p. 109028.
[27] K. Ma, Y. Fu, D. Zheng, C. Cao, X. Hu, and Y. Huang. Dynamic aggregated network for gait recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22 076–22 085.
[28] A. Sepas-Moghaddam and A. Etemad. Deep gait recognition: A survey, IEEE transactions on pattern analysis and machine intelligence, 2022, 45(1): 264–284.
[29] 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.
[30] F. Han, X. Li, J. Zhao, and F. Shen. A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition. Pattern Recognition, 2022, 125, p. 108519.
[31] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. Computer Vision and Pattern Recognition, 2019.
[32] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3d human pose and shape from a single color image. Computer Vision and Pattern Recognition, 2018.
[33] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. Computer Vision and Pattern Recognition, 2018.
[34] S. Hou, X. Liu, C. Cao, and Y. Huang. Gait quality aware network: toward the interpretability of silhouette-based gait recognition. IEEE Transactions on Neural Networks and Learning Systems, 2022.
[35] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li. Gaitgci: Generative counterfactual intervention for gait recognition, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5578–5588.
[36] N. Li and X. Zhao. A strong and robust skeleton-based gait recognition method with gait periodicity priors. IEEE Transactions on Multimedia, 2022.
[37] 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.
[38] H. Dou, P. Zhang, Y. Zhao, L. Dong, Z. Qin, and X. Li. Gaitmpl: Gait recognition with memory-augmented progressive learning. IEEE Transactions on Image Processing, 2022.
[39] X. Liu, Z. You, Y. He, S. Bi, and J. Wang. Symmetry-driven hyper feature gcn for skeleton-based gait recognition. Pattern Recognition, 2022, 125, p. 108520.
[40] S. Hou, C. Fan, C. Cao, X. Liu, and Y. Huang. A comprehensive study on the evaluation of silhouette-based gait recognition. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022.
[41] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. II-B, II-B, IV-A, IV-C
[42] C. Filipi Gonc¸alves dos Santos, D. d. S. Oliveira, L. A. Passos, R. Gonc¸alves Pires, D. Felipe Silva Santos, L. Pascotti Valem, T. P. Moreira, M. Cleison S. Santana, M. Roder, J. Paulo Papa, et al. Gait recognition based on deep learning: A survey. ACM Computing Surveys (CSUR), 2022, 55(2): 1–34.
[43] S. Hou, X. Liu, C. Cao, and Y. Huang. Set residual network for silhouette-based gait recognition. IEEE Trans. Biometrics, Behav., Identity Sci., 2021, 3(3): 384–393.
[44] H. Wu, J. Tian, Y. Fu, B. Li, and X. Li. Condition-aware comparison scheme for gait recognition. IEEE Transactions on image processing, 2021.
[45] R. Liao, Z. Li, S. S. Bhattacharyya, and G. York. Posemapgait: A model-based gait recognition method with pose estimation maps and graph convolutional networks. Neurocomputing, 2022, 501, pp. 514–528.
[46] S. Hou, C. Cao, X. Liu, and Y. Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition, 2020.
[47] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, and N. Wang. Gait recognition via disentangled representation learning. Computer Vision and Pattern Recognition, 2019.
[48] B. Lin, S. Zhang, and F. Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. ACM International Conference on Multimedia, 2020.
[49] J. N. Mogan, C. P. Lee, K. M. Lim, and K. S. Muthu. Gait-vit: Gait recognition with vision transformer. Sensors, 2022, 22(19): 7362.
[50] Q. Zou, Y. Wang, Q. Wang, Y. Zhao, and Q. Li. Deep learning-based gait recognition using smartphones in the wild. IEEE Transactions on Information Forensics and Security, 2020.
[51] Y. Wang, X. Zhang, Y. Shen, B. Du, G. Zhao, L. Cui, and H. Wen. Event-stream representation for human gaits identification using deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3436–3449.
[52] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018.
[53] C. Xu, Y. Makihara, X. Li, and Y. Yagi. Occlusion-aware human mesh model-based gait recognition. IEEE transactions on information forensics and security, 2023, 18, pp. 1309–1321.
[54] S. Choi, J. Kim, W. Kim, and C. Kim. Skeleton-based gait recognition via robust frame-level matching. IEEE Transactions on Information Forensics and Security, 2019, 14(10): 2577–2592.
[55] 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. 20249–20258.
[56] W. Yu, H. Yu, Y. Huang, and L. Wang. Generalized inter-class loss for gait recognition, in Proceedings of the 30th ACM Internationa Conference on Multimedia, 2022, pp. 141–150.
[57] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. Chinese Conferenceon Biometric Recognition, 2017.
[58] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren. End-to-end model-based gait recognition. Asian Conference on Computer Vision,

2020.

[59] X. Song, Y. Huang, C. Shan, J. Wang, and Y. Chen. Distilled light gaitset: Towards scalable gait recognition. Pattern Recognition Letters, 2022, 157, pp. 27–34.

[60] D. Das, A. Agarwal, and P. Chattopadhyay. Gait recognition from occluded sequences in surveillance sites, in European Conference on Computer Vision. Springer, 2022, pp. 703–719.

[61] S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. Computer Vision and Pattern Recognition, 2018.

[62] N. Li, X. Zhao, and C. Ma. A model-based gait recognition method based on gait graph convolutional networks and joints relationship pyramid mapping. Computer Vision and Pattern Recognition, 2020.

[63] Y. F. Song, Z. Zhang, C. Shan, and L. Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. ACM International Conference on Multimedia, 2020.

[64] H. M. Hsu, Y. Wang, C. Y. Yang, J. N. Hwang, H. L. U. Thuc, and K. J. Kim. Gaittake: Gait recognition by temporal attention and keypoint-guided embedding, in 2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022, pp. 2546–2550.

[65] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. IEEE International Conference on Computer Vision, 2020.

[66] H. Dou, P. Zhang, W. Su, Y. Yu, and X. Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition, in European Conference on Computer Vision. Springer, 2022, pp. 357–374.

[67] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. Computer Vision and Pattern Recognition, 2018.