DensePoseGait: Dense Human Pose Part-Guided for Gait Recognition

Rijun Liao, Zhu Li, *Senior Member, IEEE*, Shuvra S. Bhattacharyya, *Fellow, IEEE*,
and George York, *Senior Member, IEEE*

***Abstract*—Gait recognition is a technology that identifies human ID according to the human unique biometric gait feature. It has two popular categories, appearance-based and model-based algorithms. Appearance-based algorithms generally use human silhouettes as the initial input data. External factors such as clothing and physical carrying can drastically alter human silhouettes. In contrast, model-based algorithms tend to be more robust in regard to appearances, with human skeletons providing the initial input data in general. However, human skeletons suffer from limited information which causes an obstacle to increasing performance. In this paper, we, therefore, address this challenge by presenting two new databases, named CASIA-B-DensePose and MoBo-DensePose, which are based on the publicly available multiview database, CASIA-B and MoBo. They exploit UV coordinates of body surface and human semantic segmentation as the initial gait feature. It is less sensitive to human shape compared with human silhouettes, and has richer semantic information compared with human skeletons. In addition, we also introduce a novel model-based framework, *DensePoseGait*, to take full advantage of databases. Unlike traditional algorithms which either extract isolated local features or combine them with global features, *DensePoseGait* uses a novel way to exploit partial features. That is, human pose parts are employed as a regulator to guide the learning of global features in the training stage. Its core idea is to establish better representative features with the assistance of partial features, but not require additional calculation in the inference stage. We believe these databases and framework can offer researchers a fresh perspective on model-based gait recognition and inspire further exploration and advancements in this area.**

***Index Terms*—Gait recognition, *DensePoseGait*, part-guided learning.**

Received 20 November 2023; revised 11 March 2024 and 20 June 2024; accepted 24 June 2024. Date of publication 28 October 2024; date of current version 27 December 2024. This work was supported in part by NSF Award under Grant 2148382, and in part by AFRL SFFP at USAF Academy. This article was recommended for publication by Associate Editor R. He upon evaluation of the reviewers' comments. *(Corresponding author: Zhu Li.)*

Rijun Liao is with the School of Biomedical Engineering, Guangdong Medical University, Dongguan City 523808, Guangdong, China (e-mail: rijun.liao@gdmu.edu.cn).

Zhu Li is with the Department of Computer Science and Electrical Engineering, University of Missouri-Kansas City, Kansas City, MO 64110 USA (e-mail: zhu.li@ieee.org).

Shuvra S. Bhattacharyya is with the Department of Electrical and Computer Engineering and UMIACS, The University of Maryland at College Park, College Park, MD 20742 USA.

George York is with the Department of Electrical and Computer Engineering and UAS Research Center, U.S. Air Force Academy, Colorado Springs, CO 80840 USA.

Digital Object Identifier 10.1109/TBIOM.2024.3486732

I. Introduction

With the outbreak of the novel coronavirus 2019 (COVID-19), it has become imperative to develop biometric technologies to address various concerns arising from the rapid spread of COVID-19. Biometric technologies usually have two categories, contact and non-contact biometrics. Contact biometrics such as fingerprints and palm prints will obviously speed up the spread of the virus. For non-contact biometric technology, face recognition [1] is one of the mature biometric technologies. But identifying subjects becomes challenging when people are wearing masks. Iris recognition also faces challenges when wearing anti-virus glasses. What is more, due to the close-range collection of iris data, it also brings the risk of personnel touching the device.

Compared with the above biometrics, gait biometric has the following advantages, 1) long-distance human identification 2) no user action and cooperation required. It is particularly suitable for impeding the spread of COVID-19, monitoring people [2], video surveillance, crime prevention, and forensic identification.

There are two main approaches for gait recognition development: appearance-based and model-based algorithms. Appearance-based algorithms which generally use the human silhouettes [3], [4], [5] as initial data, as shown in Fig. 1 (b). Their advantage is that silhouettes can be easily obtained with good accuracy under simple conditions. However, they may not work well in complicated situations, like someone carrying a bag or wearing different outfits in their daily life. In contrast, model-based algorithms generally use human skeletons [6], [7], [8], [9] as initial input data, as shown in Fig. 1 (c). They are better at handling difficult scenarios because they are not affected drastically by the changes in human shape. However, the total number of human body joints is usually not more than twenty. The recognition rate is somewhat affected by the lack of information.

In this paper, instead of using human silhouettes or skeletons, we further exploit a human pose estimation feature for gait recognition, DensePose, as shown in Fig. 1 (d). DensePose is a surface-based representation of the human body. It can map all human pixels of an RGB image to the 3D surface of the human body. That is, each pixel of an RGB image has UV coordinates on the 3D surface system. It is obvious that DensePose has richer representative information compared with the human skeleton, and it is also less sensitive to human carrying or coat, even in extreme scenarios, as shown

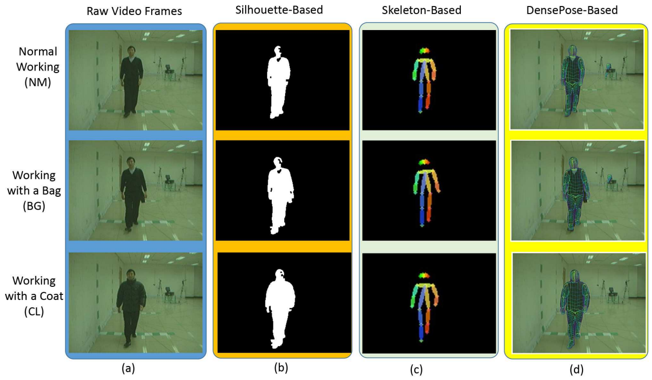


Fig. 1. Comparison of three raw input data under three walking variations. The DensePose-based input data is less sensitive to human shape compared with silhouette-based, and has richer representative features compared with skeleton-based.

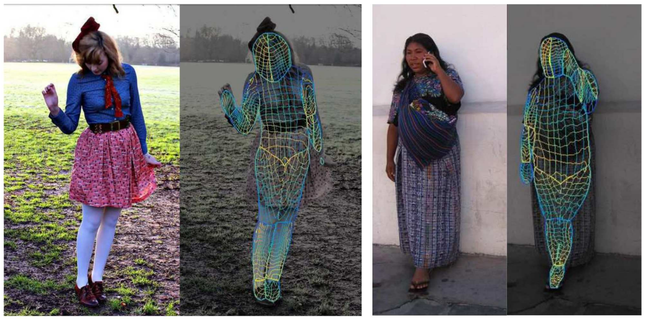


Fig. 2. DensePose [10] is robust to clothes, even with extreme skirts or dresses. It is obvious that DensePose is more robust to external factors and gait recognition utility in real applications.

in Fig. 2. Therefore, DensePose is particularly suitable as a robust gait feature.

Traditional gait recognition features predominantly fall into two categories: appearance-based and model-based algorithms. The first type method, relying on human silhouettes, is significantly affected by external factors such as clothing and physical carrying, leading to variations in the silhouette and, consequently, recognition inaccuracies. The second type method, while more robust against such appearance changes, often utilizes human skeletons that provide limited information, hindering the performance. Our motivation for employing 3D body surfaces stems from the need to address these shortcomings. The 3D body surface model offers a more holistic and stable representation of the human form, reducing sensitivity to external appearance changes and providing richer information compared to skeletal models.

The 3D body surface model captures the three-dimensional structure of the human body, offering essential view-invariant features for gait recognition. This is particularly advantageous in diverse and real-world scenarios where the subject may not always be facing the camera directly. The 3D structure provides a comprehensive and consistent dataset that is not as affected by perspective shifts, making our DensePoseGait framework robust against various viewing angles and environmental conditions.

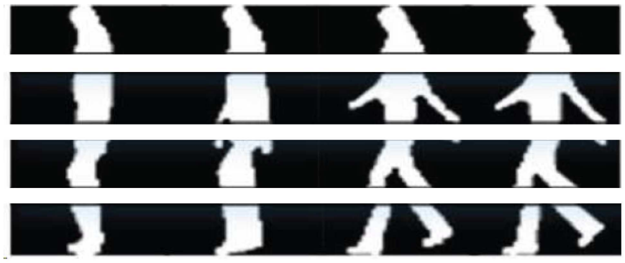


Fig. 3. General human body part division based on the fixed grid.

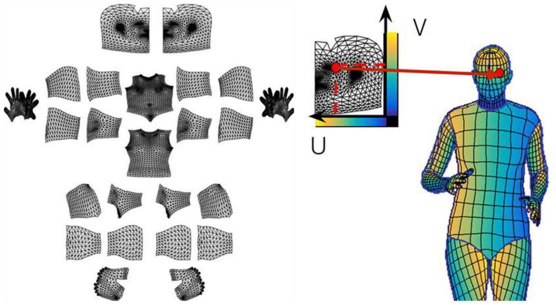


Fig. 4. Demonstration of UV coordinate maps of DensePose [10]: partitioning of the surface and correspondence to a point on a part.

Most existing gait recognition algorithms take the whole human body as a unit to extract the spatio-temporal features, thereby overlooking local part information. To further improve the gait representation ability, some works [11], [12], [13] will divide the human body into several parts, and extract local features or combine them with global features for final identification. Such as Fan et al. [11] divide the human body into 4 parts from up to down, and extracts spatio-temporal representations from each body part. And Hossain et al. [12] divide the human body into 8 parts to solve the problem of clothing. However, those dividing are typically based on the fixed grid, and can not exactly segment the human semantic body, as shown in Fig. 3. It would totally have a negative influence on the modeling of the human body local movement.

In order to address the above problem, we further exploit the partial information of DensePose, which can accurately divide the human body into 24 parts, as shown in Fig. 4. Unlike the above methods [11], [12], [13] which focus on establishing partial features or combining them with global features, we, introduce a novel model-based gait recognition framework, *DensePoseGait*. *DensePoseGait* framework employs pose parts to guide the representative feature learning in the training stage. This allows us to create higher-quality representative features without requiring additional calculations during the inference stage. The idea of *DensePoseGait* is inspired by one work [14] of IEEE International Conference on Computer Vision Workshops, which verifies that part-pose guided feature learning is beneficial to re-identification. We extend this network on the task of gait recognition and exploit a sequence of dense pose maps rather than just a single-frame skeleton image.

In summary, our major contributions are:

- This paper presents two new databases, named CASIA-B-DensePose, Mobo-DensePose, which bring a new way to exploit model-based gait recognition for researchers. DensePose databases are based on the publicly available multiview databases, CASIA-B and Mobo. It exploits UV coordinates of the human surface and human semantic segmentation as the initial gait feature. It is less sensitive to human shape compared with human silhouettes, and has richer information compared with human skeletons.
- To make the most of DensePose gait databases, we introduce a novel model-based gait recognition framework, called *DensePoseGait*, it carries richer information and accurate body segmentation compared with human skeletons. In addition, some existing algorithms leverage partial features to enhance individual identification, but they lack highly accurate semantic segmentation. With *DensePoseGait* framework, this issue is resolved since it enables accurate 24-part semantic segmentation of the human body.
- Unlike traditional algorithms which either extract isolated local features or combine them with global features, we use accurate semantic segmented parts as a regulator to make an alignment constraint on global gait feature learning. The purpose is to create better representative features by making use of partial features, but without requiring extra calculations in the inference stage.

II. Related Work

Gait recognition algorithms can be roughly divided into two categories: appearance-based and model-based. In this section, we will give a quick overview of the existing algorithms in the field. We also will review DensePose algorithm in brief.

A. Model-Based Algorithms

Model-based algorithms extract features through the modeling of the human body structure and the examination of movement patterns of various body parts. Unlike appearance-based approaches, which are highly sensitive to human appearance, these methods prove to be more robust to variations due to their focus on the analysis of movement.

Studies [15], [16], [17] conducted in the early stage of research indicate that human body movement patterns have the potential to reveal human identity. As advancements in pose estimation algorithms continue, this concept is becoming increasingly relevant. Some researchers [6], [18], [19] improve the performance of model-based methods greatly with the help of pose skeleton estimation algorithms. In 2017, Liao et al. [6] extract 2D human pose skeletons and put into PTSN [6] network. In 2018, 3D pose skeletons are used in PTSN-3D [7] network for gait recognition. In 2020, PoseGait [9] has made a remarkable debut and achieved a commendable recognition rate by deep analysis of the 3D pose skeleton. In the same year, OU-ISIR create a gait dataset with pose sequence [8] for public research. In 2022, PoseMapGait [13] further improve the gait performance by using pose heatmaps rather than relying on skeleton joints coordinates.

To address the limitations associated with skeletal representations, recent studies such as those by Fan et al. [20] and Liao et al. [13] have employed skeleton-maps to augment skeletal information, thereby enhancing gait robustness to some extent. Skeleton-maps entail the prediction of joint maps surrounding human body joints. In contrast, dense pose features predict UV coordinate maps across the human surface, indicating that dense pose representations offer richer information compared to skeleton-maps.

Furthermore, the advancement in human parsing techniques has found successful applications in gait recognition, as demonstrated in works by Wang et al. [21] and Zheng et al. [22]. Human parsing involves segmenting the human body into various parts, thereby enhancing pose features. In contrast, dense pose features not only encompass attributes of human segmentation similar to parsing but also estimate UV coordinates for every human joint on the surface. Thus, while human parsing primarily focuses on human body segmentation, dense pose extends its focus to include human surface modeling, providing a more comprehensive representation of the human form.

Model-based approaches have seen substantial growth, thanks to the above works. The recognition rate still requires further enhancement due to limited skeleton information. In contrast, the input data of the proposed *DensePoseGait* framework has richer dense human pose feature representative information compared with skeletons, skeleton maps, or parsing.

B. Appearance-Based Algorithms

Appearance-based algorithms generally extract features from the human silhouettes. Some algorithms would create a gait template from a sequence of human silhouettes, e.g., Gait Energy Image (GEI) template [23] and Chrono-Gait Image (CGI) template [24]. Template-based algorithms [3], [4], [25] don't require much computational cost, but they usually miss mass temporal information.

In order to extract more temporal information, some algorithms [11], [26], [27] apply a deep convolutional neural network extract representative feature from a sequence of human silhouettes. Chen et al. propose pyramid attention [28] and GaitAMR [29], and Dou et al. [30] introduce GaitMPL framework, they have achieved high accuracy on the gait recognition.

Appearance-based algorithms can perform well in some conditions, like view variation. However, they have a serious shortcoming, that is, human silhouettes would be changed dramatically under difficult scenarios, like the condition of wearing a big cloth, as shown in Fig. 1 (b). By contrast, the input data of our proposed *DensePoseGait* framework doesn't have this shortcoming, and also have accurate semantic segmentation of body parts.

C. DensePose

Unlike traditional human pose estimation which evaluates human skeleton joint coordinates, dense pose estimation [10]

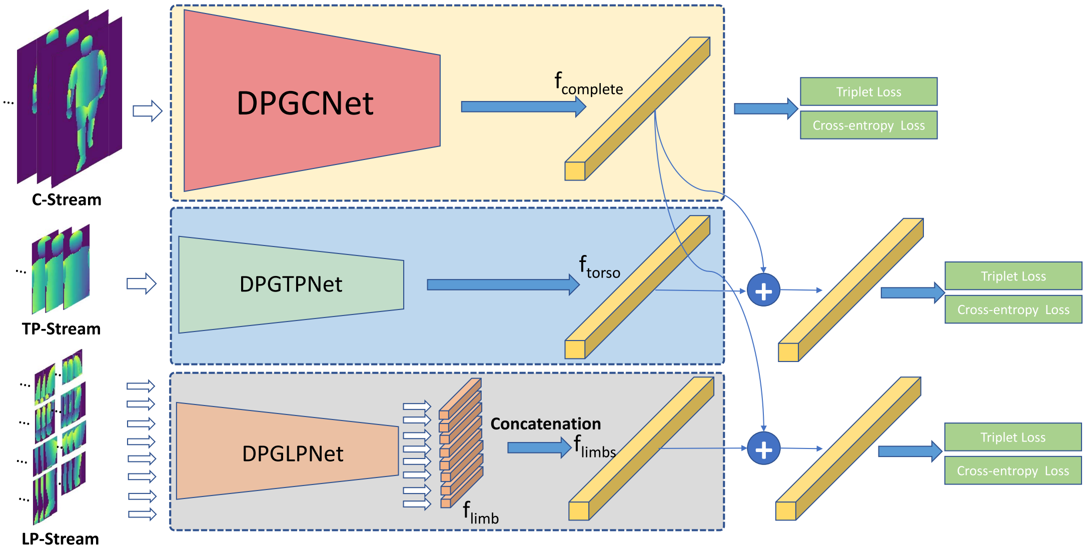


Fig. 5. The overview of the proposed framework *DensePoseGait*. Given a video of people's walking sequence. DensePose [10] will extract UV coordinate maps and corresponding human body parts. Then, UV coordinate maps will be put into one C-stream and two partial streams. C-stream is the mainstream for extracting global features from complete gait images. In terms of two partial streams, TP-stream and LP-stream, are used to extract local features from torso and limb gait patches, respectively. In the training phase, local features play a regulating role in global feature learning. In the inference phase, only the C-stream is used for gait recognition.

aims at mapping all human pixels of an RGB image to the 3D surface of the human body. DensePose will semantically divide the human body into 24 parts, including Head, Torso, Lower/Upper Arms, Lower/Upper Legs, Hands, and Feet. Each part will estimate UV coordinate maps on the surface-based representations of the human body. For every pixel, will determine which surface part it belongs to, and where on the 2D paremeterization of the part it corresponds to, as shown in Fig. 4. The yellow color pixels have a higher coordinate value than the green color pixels in the UV surface system.

III. Proposed Method

In this section, we will present the pipeline of *DensePoseGait* framework, as shown in Fig. 5. *DensePoseGait*. Given a video of people's walking sequence. DensePose [10] will be used to extract UV coordinate maps and corresponding human body parts. In order to make full use of local movement patterns and global movement patterns, UV coordinate maps will be put into three streams: one C-stream and two partial streams. C-stream is the mainstream for extracting global features from complete gait images. In terms of two partial streams, TP-stream and LP-stream, are used to extract local features from torso and limb gait patches, respectively. In the training phase, local features play a regulating role in global feature learning. In the inference phase, only the C-stream is used for gait recognition.

A. DensePose UV Coordinate Maps

In this section, we will describe the generation of robust gait input features from the dense pose. Given a sequence of people walking images, DensePose [10] will extract UV coordinate maps and corresponding human body parts. As shown in Fig. 6, each frame will generate three maps, U coordinate maps, V coordinate maps, and human body part index maps.

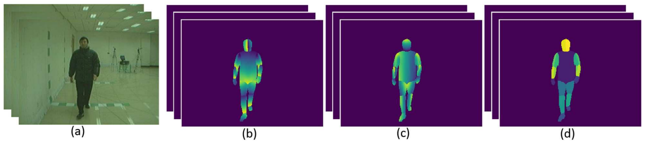


Fig. 6. Demonstration of output result of DensePose [10]. a) Input image. b) U coordinate maps. c) V coordinate maps. d) human body part index maps.

UV coordinates maps are surface-based representations of the human body, the detail is described in Section II-C. Each map will be aligned based on methods in [31].

In order to make full use of local movement patterns and global movement patterns, rather than directly use dense poses directly, we generate three types of maps as input features, as shown in Fig. 5. Two part maps will guide the learning of global features.

Complete Body Maps: It consists of U and I two coordinates maps. The sequence of complete body maps will be put into the Dense Pose Gait Complete Network (DPGCNet) to model the pattern of the global movement. Given a video with N frames, a sequence of complete body maps is formulated as:

$$C = \{c_i | i = 1, 2, \ldots, N\} \tag{1}$$

where c_i is the complete body maps at current time i, including UI coordinates maps.

Torso Part Maps: It also consists of U and I two coordinates maps, but only will the torso part according to the human body part index maps. The sequence of torso part maps will be put into the Dense Pose Gait Torso Part Network (DPGTPNet) to extract the pattern of the body truck. Given a video with N frames, a sequence of torso part maps is formulated as:

$$T = \{t_i | i = 1, 2, \ldots, N\} \tag{2}$$

where t_i is the torso part maps at current time i, as shown in the middle input image of DPGTPNet in Figure 5. t_i includes the head and human torso two parts, it is extracted from human UV coordinate maps according to human body part index maps, as shown in Figure 6. The human body part index maps has 24 parts, including Head, Torso, Lower/Upper Arms, Lower/Upper Legs, Hands, and Feet. The index range is from 0 to 23.

Limb Part Maps: It consists of 8 parts: left upper arm, right upper arm, left lower arm + hand, right lower arm + hand, left upper leg, right upper leg, left lower leg + feet, and right lower leg + feet. And each part contains U and I two coordinates maps. The sequence of limb part maps will be put into the Dense Pose Gait Limbs Part Network (DPGLPNet) to extract the movement pattern from each limb part. Given a video with N frames, a sequence of limb part maps is formulated as:

$$L = \{l_i | i = 1, 2, \ldots, N\} \tag{3}$$

where l_i is the torso part maps at current time i, as shown in the third input image of DPGLPNet in Figure 5. l_i consists of 8 human limb parts, it is extracted from human UV coordinate maps according to human body part index maps, as shown in Figure 6.

B. DensePoseGait Framework

The purpose of the DensePoseGait network is to leverage part-pose representations for enhancing the learning of global gait features. The network comprises three integral streams: the Complete Body Stream (C-Stream), the Torso Part Stream (TP-Stream), and the Limbs Part Stream (LP-Stream), as now clearly illustrated in the revised Fig. 5. The backbone network of three streams are all based on the GaitSet [27], as shown in Fig. 7

C-Stream (Complete Body Stream): The C-Stream is responsible for capturing and analyzing the global gait features. It serves as the backbone of the network, synthesizing the overall gait pattern from a holistic perspective.

TP-Stream (Torso Part Stream): The TP-Stream focuses specifically on the torso part, extracting localized features that are crucial for understanding the upper body movements and postures in gait analysis.

LP-Stream (Limbs Part Stream): Similarly, the LP-Stream is dedicated to the limbs, analyzing lower body movements and contributing to the understanding of the gait dynamics.

The feature $f_{complete}$, supervised jointly by the C-Stream, TP-Stream, and LP-Stream, embodies both the global and local aspects of gait features. Specifically, f_{ct} and f_{cl} represent the global feature (extracted from the C-Stream) and the local feature (from the Part-Streams), respectively. The optimization of $f_{complete}$ loss is a collaborative effort of all three streams, ensuring a comprehensive and nuanced feature development.

During gradient backpropagation in the training phase, the C-Stream's parameters are adjusted in response to the gradient loss from the fused features of the TP-Stream and LP-Stream. This process ensures that the C-Stream is continuously refined by the influence of local features, leading to a more accurate and holistic gait representation. Consequently, the TP-Stream and LP-Stream play a crucial regularization role in guiding the feature learning for the C-Stream.

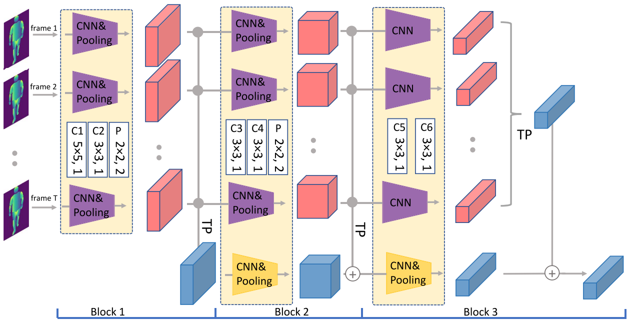


Fig. 7. Backbone Network. Spatial features (red cube) are extracted from each frame by traditional CNN. Temporal features (blue cube) are extracted by temporal pooling (TP).

Complete Body Stream (C-Stream): This stream is mainly to extract rich global spatial and temporal feature representations. A sequence of global complete body maps C are put into the Dense Pose Gait Complete Network (DPGCNet) to extract global feature vector $f_{complete}$. DPGCNet uses GaitSet [27] network as backbone, as shown in Fig. 7. It consists of three blocks, and each block will extract spatial and temporal features, formulated as follows:

$$f_{complete} = G(F(C)) \tag{4}$$

where F is a traditional convolutional network which can extract the spatial features ($f_{complete} = \{f^i | i = 1, 2, \ldots, N\}$) from each complete body maps ($C = \{c_i | i = 1, 2, \ldots, N\}$). G is used to extract temporal features from spatial features. A sequence of spatial features ($\{f^i | i = 1, 2, \ldots, N\}$) would put into G function. G is achieved by an operation called temporal pooling (TP), $G(\cdot) = max(\cdot) + mean(\cdot) + median(\cdot)$, which aims to aggregate gait information of elements in the time sequence. Other than those used in traditional convolutional neural networks, it preserves strong temporal information, as well as sufficiently spans spatial information. The diagram of temporal pooling can be shown in Fig. 7, $f_{complete}$ is the output features from complete body maps.

The essence of TP is to capture different statistical aspects of the spatial features across the temporal dimension. Specifically, the max operation is used to capture the most prominent feature in the sequence, which often corresponds to the most significant movement in a gait cycle. The mean operation provides an average representation of the features over time, effectively smoothing out anomalies and highlighting consistent patterns. Lastly, the median operation contributes by pinpointing the central tendency of the features, which is crucial in identifying typical postures or movements in the gait cycle.

By combining these three operations, TP effectively synthesizes a comprehensive temporal profile from the spatial features. This aggregation method preserves temporal information by encapsulating the variability (max), consistency (mean), and typicality (median) of the gait patterns over time. Therefore, while each individual operation (max, mean, median) contributes a unique perspective on the spatial features, their summation in TP provides a multi-faceted temporal representation. This representation is crucial for accurate gait analysis, as it ensures that the model is not only informed by the most extreme or average patterns but also by the typical and consistent features that characterize an individual's gait over time.

Torso Part Stream (TP-Stream): TP-Stream focuses on extracting spatio-temporal representations from a sequence of torso part maps. This is because the torso part of a human has the richest stature information that can identify the individual person. T-Stream consists of the Dense Pose Gait Torso Part Network (DPGTPNet), which includes block 1 and block 2 of the backbone network (Fig. 7). The output f_{torso} of DPGTPNet is formulated as:

$$f_{torso} = G(F(T)) \tag{5}$$

Limbs Part Stream (LP-Stream): To learn the local details of individual parts area rather than mixing them all together, the Dense Pose Gait Limbs Part Network (DPGLPNet) with multiple branches are used to learn the limbs feature map. First, we use the 8 parallel independent DPGLPNet networks to learn the local features $f_{limb,l}$, $l = 1, 2, 3, 4, 5, 6, 7, 8$ from 8 limb part maps $T = \{t_i | i = 1, 2, \dots, N\}$. And then concatenate 8 local features $f_{limb,l}$ into one limb representative features f_{limbs}, formulated as equation (6). Considering the size of limbs is smaller than the size of the complete body, the DPGLPNet network only consists of block 1 of the backbone network (Fig. 7).

$$f_{limbs} = G(F(L)) \tag{6}$$

C. Feature Fusion

To improve the learning and alignment of global features using local part features, we fuse the part and global features from the three streams by adding corresponding feature vectors:

$$f_{ct} = f_{complete} + f_{torso} \tag{7}$$

$$f_{cl} = f_{complete} + f_{limbs} \tag{8}$$

Here, $f_{complete}$ represents the global feature extracted by the C-Stream. The local features f_{torso} and f_{limbs} are extracted by the TP-Stream and LP-Stream, respectively. The fused feature f_{ct} combines the global feature with the torso feature, while f_{cl} combines the global feature with the limbs feature.

During training, these fused features help adjust the learning of the global feature $f_{complete}$. By integrating local features (from the TP-Stream and LP-Stream) with the global feature, the learning process for $f_{complete}$ is guided by detailed local information.

D. Loss Function

To optimize the three streams (C-Stream, TP-Stream, and LP-Stream), we employ a combination of cross-entropy loss and triplet loss. The cross-entropy loss is used for gait ID identification, while the triplet loss enhances inter-class variation and reduces intra-class variation. The total loss is formulated as a weighted sum of these two losses:

$$L_{total} = \alpha L_{cross-entropy} + \beta L_{triplet} \tag{9}$$

where α and β are the weights for the cross-entropy loss and triplet loss, respectively. In our experiments, we set $\alpha = \beta = 1$.

The feature $f_{complete}$, which represents the global gait pattern, is supervised by the C-Stream, TP-Stream, and LP-Stream. The fused features f_{ct} (from C-Stream and TP-Stream) and f_{cl} (from C-Stream and LP-Stream) consist of both global and local information. The learning process for $f_{complete}$ is influenced by these fused features, ensuring that local features guide the global feature learning.

During backpropagation, the gradient of the total loss L_{total} with respect to the parameters of the C-Stream (θ_C) is influenced by the losses from all three streams:

$$\frac{\partial L_{total}}{\partial \theta_C} = \frac{\partial L_C}{\partial \theta_C} + \frac{\partial L_{TP}}{\partial \theta_C} + \frac{\partial L_{LP}}{\partial \theta_C} \tag{10}$$

Here, $\frac{\partial L_C}{\partial \theta_C}$ is the gradient from the C-Stream, $\frac{\partial L_{TP}}{\partial \theta_C}$ is the gradient from the TP-Stream, and $\frac{\partial L_{LP}}{\partial \theta_C}$ is the gradient from the LP-Stream. This equation shows that the parameter updates for the C-Stream are influenced by both the global feature learning and the detailed local features from the TP-Stream and LP-Stream.

By integrating local features (f_{torso} and f_{limbs}) into the global feature learning process ($f_{complete}$), the C-Stream can adjust its parameters in response to the detailed local insights. This ensures that the C-Stream's learning is continually refined and guided by the local information, resulting in a more comprehensive and nuanced representation of gait patterns.

In summary, the combination of local and global features during training allows the C-Stream to benefit from the detailed features captured by the TP-Stream and LP-Stream, leading to a robust and thorough learning process that enhances the model's capability for gait identification.

IV. Experiments

A. Datasets

CASIA-B [32]: It is a widely applied gait dataset. It has 124 subjects. Each subject has 10 sequences, including 6 sequences of normal walking (NM), 2 sequences of walking with a bag (BG), and 2 sequences of walking with a coat (CL), as shown in Fig. 1 (a). What's more, each sequence has 11 views {0°,18°,$\cdots$, 180°} with around 80 frames. The number of total frames is around $124 \times 10 \times 11 \times 80 = 1,091,200$ images. The experimental setting of training and testing, as shown in Table I.

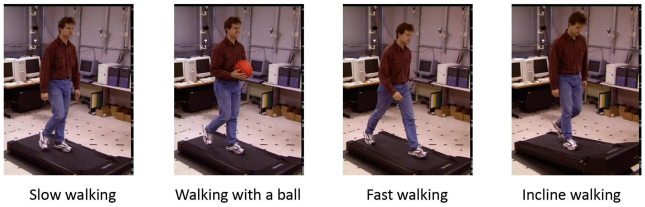
OU-MVLP [31], **OU-ISIR [34], Gait3D [35] and GREW [36]:** They are also very popular data with large subjects. We plan to do some experiments on these datasets. However, the original RGB videos are not open to the public due to the privacy issue. It is hard to extract dense pose maps without RGB videos. Therefore, we don't perform experiments on these datasets.

TABLE I
EXPERIMENTAL SETTING ON CASIA-B DATASET. NM: NORMAL WALKING, BG: WALKING WITH A BAG, CL: WALKING WITH A COAT

Training	Testing	
	Gallery Set	Probe Set
ID: 001-062 Seqs: NM01-NM06 BG01-BG02, CL01-CL02	ID: 063-124 Seqs: NM01-NM04	ID: 063-124 Seqs: NM05-NM06 BG01-BG02, CL01-CL02

TABLE II
EXPERIMENTAL SETTING ON THE CMU MOTION OF BODY (MOBO) DATASET [33]

Training	Test	
	Gallery Set	Probe Set
ID: 01-13 slow walking, fast walking, incline walking, walking with a ball	ID: 14-25 slow walking	ID: 14-25 fast walking, incline walking, walking with a ball



Slow walking Walking with a ball Fast walking Incline walking

Fig. 8. Four walking variations on MoBo [33] dataset.

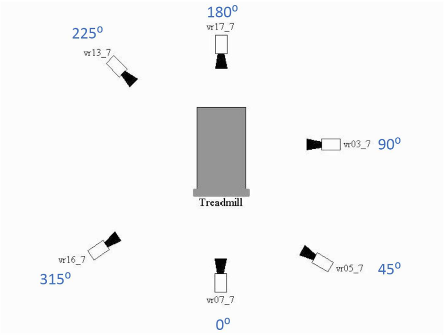


Fig. 9. View angle definition on MoBo [33] dataset.

The CMU Motion of Body (MoBo) dataset [33]: As mention above, OU-MVLP [31] and OU-ISIR [34] are not available for the experiment. Therefore, we perform experiments on the MoBo [33] dataset. It has 25 subjects, and each subject has 4 conditions, that is, slow walking, fast walking, incline walking, and walking with a ball, as shown in Fig. 8. In addition, each subject is captured using 6 cameras distributed evenly around the treadmill, cameras are defined as *vr*03_7, *vr*05_7, *vr*07_7, *vr*13_7, *vr*16_7, and *vr*17_7. Similar to the angle definition of the CASIA-B dataset, we define the angle set of MoBo dataset is {0°, 45°, 90°, 180°, 225°, 315°}, as shown in Fig. 9. The experimental setting of training and testing of MoBo [33] dataset, as shown in Table II.

B. Comparison With Model-Based Algorithms on CASIA-B Dataset

We compare our proposed method *DensePoseGait* with recent state-of-the-art model-based algorithms on CASIA-B dataset. Including methods based on the 2D human skeleton, PTSN [6], methods based on the 3D human skeleton, PTSN-3D [7] and PoseGait [9], and methods based on the 2D human pose heatmap, PoseMapGait [13]. The performance is shown in Table III.

From Table III, we can see that *DensePoseGait* can achieve the highest performance under the three walking conditions, that is, 77.5% (NM), 65.2% (BG), and 45.2% (CL), respectively. The gap of mean accuracy between the *DensePoseGait* (58.1%) and the state-of-the-art method PoseMapGait (65.2%) can even reach 7.1% on the variation of carrying a bag. The input features of compared methods [6], [7], [9], [13] are all based on human skeletons. Unlike these approaches which considered several human joint coordinates modeling, we not only generate rich body UV coordinates (dense pose maps) as gait features, but also make full use of body parts to promote the learning of global features. The comparison shows that dense pose maps can further improve the development of model-based approaches for gait recognition.

C. Comparison With Appearance-Based Algorithms on CASIA-B Dataset

We also compare *DensePoseGait* with recent appearance-based algorithms. Including SPAE [37], GaitGAN [38], GaitGANv2 [25], DV-GEIs-pre [4], and DV-GEIs [3]. The initial input features of those methods are all based on human silhouettes or their variants. The experimental results can be shown in Table IV.

From Table IV, we can see that our proposed method gets better performance compared with these silhouette-based methods [3], [4], [25], [37], [38]. There are obvious gaps in the conditions of carrying a bag or wearing a coat. It is obvious shows that the input data of *DensePoseGait* is less sensitive to human shape compared with the human silhouette.

We also compare our method with contemporary appearance-based methods, including GaitSet [27], GaitPart [11], 3DLocal [39], and CSTL [40]. In our study, we delineated three distinct experimental scenarios to evaluate our models. For the initial scenario, we allocated the first set of 62 subjects to the training group, with the subsequent 62 subjects serving as the test group. The second scenario

TABLE III
AVERAGE RECOGNITION RATE (%) COMPARISONS WITH MODEL-BASED APPROACHES ON CASIA-B DATASET. EXCLUDING IDENTICAL-VIEW CASES. (NM: NORMAL WALKING, BG: WALKING WITH A BAG, CL: WALKING WITH A COAT)

Gallery angle NM #1-4	0°-180°											
Probe angle NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [6]	34.5	45.6	49.6	51.3	52.7	52.3	53	50.8	52.2	48.3	31.4	47.4
PTSN-3D [7]	38.7	50.2	55.9	56	56.7	54.6	54.8	56	54.1	52.4	40.2	51.9
PoseGait [9]	48.5	62.7	66.6	66.2	61.9	59.8	63.6	65.7	66	58	46.5	60.5
PoseMapGait [13]	59.9	76.2	81.7	83.1	76.8	76.1	76.3	81.1	79.6	75.4	66.1	75.7
DensePoseGait (ours)	65.7	79.7	82.8	84.4	79.4	77.9	80.1	83.4	83.7	74.3	61.5	**77.5**
Probe angle BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [6]	22.4	29.8	29.6	29.2	32.5	31.5	32.1	31	27.3	28.1	18.2	28.3
PTSN-3D [7]	27.7	32.7	37.4	35	37.1	37.5	37.7	36.9	33.8	31.8	27	34.1
PoseGait [9]	29.1	39.8	46.5	46.8	42.7	42.2	42.7	42.2	42.3	35.2	26.7	39.6
PoseMapGait [13]	47.7	56.1	63.9	63.3	64.2	59.5	58.1	61.5	61.9	58.2	44.3	58.1
DensePoseGait (ours)	55.4	70.4	76.6	73.3	65.6	65.3	68.1	71.0	69.8	57.3	44.8	**65.2**
Probe angle CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
PTSN [6]	14.2	17.1	17.6	19.3	19.5	20	20.1	17.3	16.5	18.1	14	17.6
PTSN-3D [7]	15.8	17.2	19.9	20	22.3	24.3	28.1	23.8	20.9	23	17	21.1
PoseGait [9]	21.3	28.2	34.7	33.8	33.8	34.9	31	31	32.7	26.3	19.7	29.8
PoseMapGait [13]	30.4	41.9	45.2	48.9	47.3	48.1	46.5	44.9	36.0	34.5	29.6	41.2
DensePoseGait (ours)	41.8	47.7	49.7	50.3	46.5	46.0	49.5	47.8	47.4	39.4	29.3	**45.2**

TABLE IV
COMPARISONS WITH APPEARANCE-BASED ALGORITHMS AT AVERAGE ACCURACY (%) ON CASIA-B DATASET. EXCLUDING IDENTICAL-VIEW CASES. (NM: NORMAL WALKING, BG: WALKING WITH A BAG, CL: WALKING WITH A COAT)

Gallery angle NM #1-4	0°-180°											
Probe angle NM #5-6	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [37]	50.0	58.1	61.0	63.3	64.0	62.1	62.3	66.3	64.4	54.5	46.7	59.3
GaitGAN [38]	41.9	53.5	63.0	64.5	63.1	58.1	61.7	65.7	62.7	54.1	40.6	57.2
GaitGANv2 [25]	48.1	61.9	68.7	71.7	66.7	64.8	66.0	70.2	71.6	58.9	46.1	63.1
DV-GEIs-pre [4]	64.5	76.2	81.3	80.8	77.1	72.6	74.4	78.9	80.6	75.6	63.7	75.1
DV-GEIs [3]	63.1	79.4	84.6	79.8	77.0	72.6	77.4	80.3	84.0	78.5	63.7	76.4
DensePoseGait (ours)	65.7	79.7	82.8	84.4	79.4	77.9	80.1	83.4	83.7	74.3	61.5	**77.5**
Probe angle BG #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [37]	34.0	38.6	42.1	42.7	39.0	32.8	31.3	39.9	41.0	35.7	32.3	37.2
GaitGAN [38]	28.5	35.2	42.7	34.4	38.0	33.5	36.2	44.8	41.8	33.3	23.6	35.6
GaitGANv2 [25]	37.2	43.4	46.6	46.0	47.6	41.5	41.2	48.5	48.8	42.2	31.6	43.1
DV-GEIs [3]	47.5	59.6	64.2	66.3	61.3	56.7	63.4	63.3	61.8	57.5	47.0	59.0
DensePoseGait (ours)	55.4	70.4	76.6	73.3	65.6	65.3	68.1	71.0	69.8	57.3	44.8	**65.2**
Probe angle CL #1-2	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
SPAE [37]	21.5	25.4	27.3	28.1	26.9	22.2	22.3	26.3	24.8	21.5	19.6	24.2
GaitGAN [38]	9.8	15.2	24.8	25.0	24.7	19.9	22.7	24.5	27.7	18.0	11.9	20.4
GaitGANv2 [25]	20.7	23.1	26.6	30.8	28.2	23.0	24.4	27.4	24.2	21.9	16.0	24.2
DV-GEIs [3]	30.2	43.3	43.4	43.1	43.6	41.9	40.0	40.3	41.4	38.7	29.9	39.6
DensePoseGait (ours)	41.8	47.7	49.7	50.3	46.5	46.0	49.5	47.8	47.4	39.4	29.3	**45.2**

was structured with an augmented training group comprising the initial 74 subjects, leaving a smaller cohort of 50 subjects for testing. The third and final scenario further expanded the training group to encompass the first 100 subjects. As shown in Table V.

These established methods have set a high bar for gait recognition accuracy, particularly under variable conditions such as wearing coats or carrying bags. Their strategies revolve around refining the human silhouette's representation. However, the silhouette is inherently limited by its inability to abstract away clothing and accessory variations.

In contrast, *DensePoseGait* introduces a paradigm shift by theoretically eliminating these variations at the source. Instead of iterating upon silhouette processing, it leverages a novel feature type that transcends the conventional silhouette. Initial results suggest that while *DensePoseGait*‘s accuracy currently trails that of its appearance-based counterparts–partially due to the 3D model's occasional omission of fine body details–the method's growth rates, as demonstrated in our latest dataset expansions, indicate a robust learning curve and substantial adaptability.

DensePoseGait, as substantiated by our experimental data, shows a remarkable capacity for improvement and adaptability. In the normal walking condition (NM), our method improved from 77.5% to 93.5% in accuracy as the training set expanded from 62 to 100 subjects, a growth rate of 17.1%. This is significantly higher than the other methods, which show more modest improvements over the same interval. For example, GaitSet [27], a leading method, shows a growth rate of 4.0%, from 92.0% to 95.815%.

Moreover, under the more challenging conditions where subjects carried a bag (BG), *DensePoseGait*‘s growth rate was an impressive 22.0%, and in the coat-wearing scenario (CL), it demonstrated a staggering growth rate of 42.2%, highlighting

TABLE V
COMPARISON ON CASIA-B DATASET. EXCLUDING IDENTICAL-VIEW CASES

Conditions	Method Types	Methods	Training Subjects			Growth Rate		Gap with *DensePoseGait*		
			62	74	100	62 to 74	62 to 100	62	74	100
NM	Appearance-based	GaitSet [27]	92.0	95.0	95.815	3.3%	4.0%	14.5	9.5	2.3
		GaitPart [11]	92.4	96.2	97.4	4.1%	5.1%	14.9	10.7	3.9
		3DLocal [39]	94.0	97.5	98.5	3.7%	4.6%	16.5	12.0	5.0
		CSTL [40]	94.5	97.8	98.9	3.5%	4.4%	17.0	12.3	5.4
	Model-based	*DensePoseGait* (ours)	77.5	85.5	93.5	10.3%	17.1%	-	-	-
BG	Appearance-based	GaitSet [27]	84.3	87.2	91.78	3.4%	8.1%	19.1	13.8	8.2
		GaitPart [11]	87.9	91.5	94.5	4.1%	7.0%	22.7	18.1	10.9
		3DLocal [39]	90.9	94.3	96.6	3.7%	5.9%	25.7	20.9	13.0
		CSTL [40]	90.1	93.6	96.1	3.9%	6.2%	24.9	20.2	12.5
	Model-based	*DensePoseGait* (ours)	65.2	73.4	83.6	12.6%	22.0%	-	-	-
CL	Appearance-based	GaitSet [27]	62.5	70.4	83.144	12.6%	24.8%	17.3	8.0	4.9
		GaitPart [11]	70.7	78.7	84.7	11.3%	16.5%	25.5	16.3	6.5
		3DLocal [39]	75.5	83.7	89.9	10.9%	16.0%	30.3	21.3	11.7
		CSTL [40]	75.8	84.2	90.3	11.1%	16.1%	30.6	21.8	12.1
	Model-based	*DensePoseGait* (ours)	45.2	62.4	78.2	38.1%	42.2%	-	-	-

the method's exceptional potential for dealing with external variations. The performance increment is noteworthy, especially considering that in the CL condition, *DensePoseGait*'s initial accuracy of 45.2% rose significantly to 78.2%, while other methods, although starting from a higher baseline, showed lesser improvements.

In Table V, we have conducted an analysis to quantify the accuracy disparity between traditional appearance-based algorithms and our model-based algorithm *DensePoseGait*. The results clearly illustrate a narrowing accuracy gap between *DensePoseGait* and other methods as the size of the training dataset increases. This trend is particularly noteworthy, as seen in the transition from a baseline dataset of 62 subjects to a more extensive dataset encompassing 100 subjects.

Compared to established appearance-based algorithms such as GaitSet [27], the gap in performance diminishes notably under various conditions on the conditions of normal condition (2.3) and wearing a coat (4.9). Although *DensePoseGait* may initially lag behind due to the dense pose representation potentially missing some fine-grained body details compared to traditional human silhouettes, its trajectory indicates rapid improvement and adaptability as it learns from a larger pool of training data.

The tables above illustrate the cross-view capabilities of gait recognition. Notably, the *DensePoseGait* framework demonstrates its optimal performance under the identity-view condition. As depicted in Fig. 11, when the probe angle matches the gallery angle, DensePoseGait achieves remarkable accuracy rates, reaching nearly 100%, 95%, and 85% for the NM, BG, and CL conditions, respectively, even with only 62 training subjects. This highlights the robustness of dense pose gait features, indicating their lower sensitivity to variations in human shape and resilience to factors like carrying bags or wearing different clothing styles.

The observed trend suggests that the distinctive feature set of *DensePoseGait*, which theoretically mitigates variations that conventional silhouettes struggle to capture, exhibits high scalability and shows significant performance enhancement with the provision of additional data. This scalability is imperative for real-world applications where environmental conditions can vary widely.

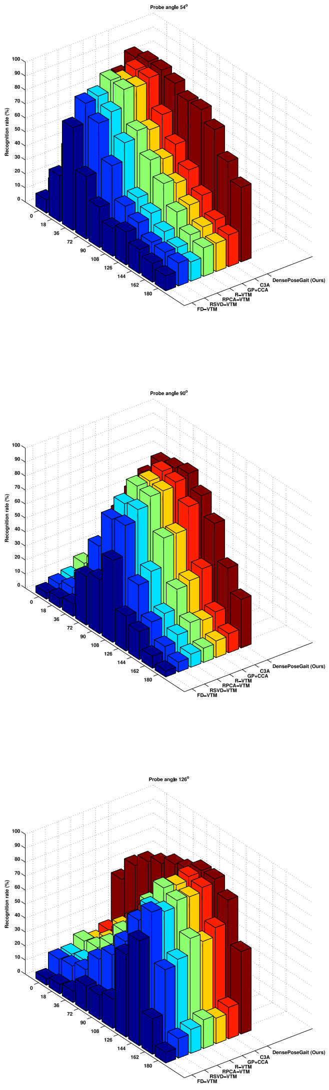
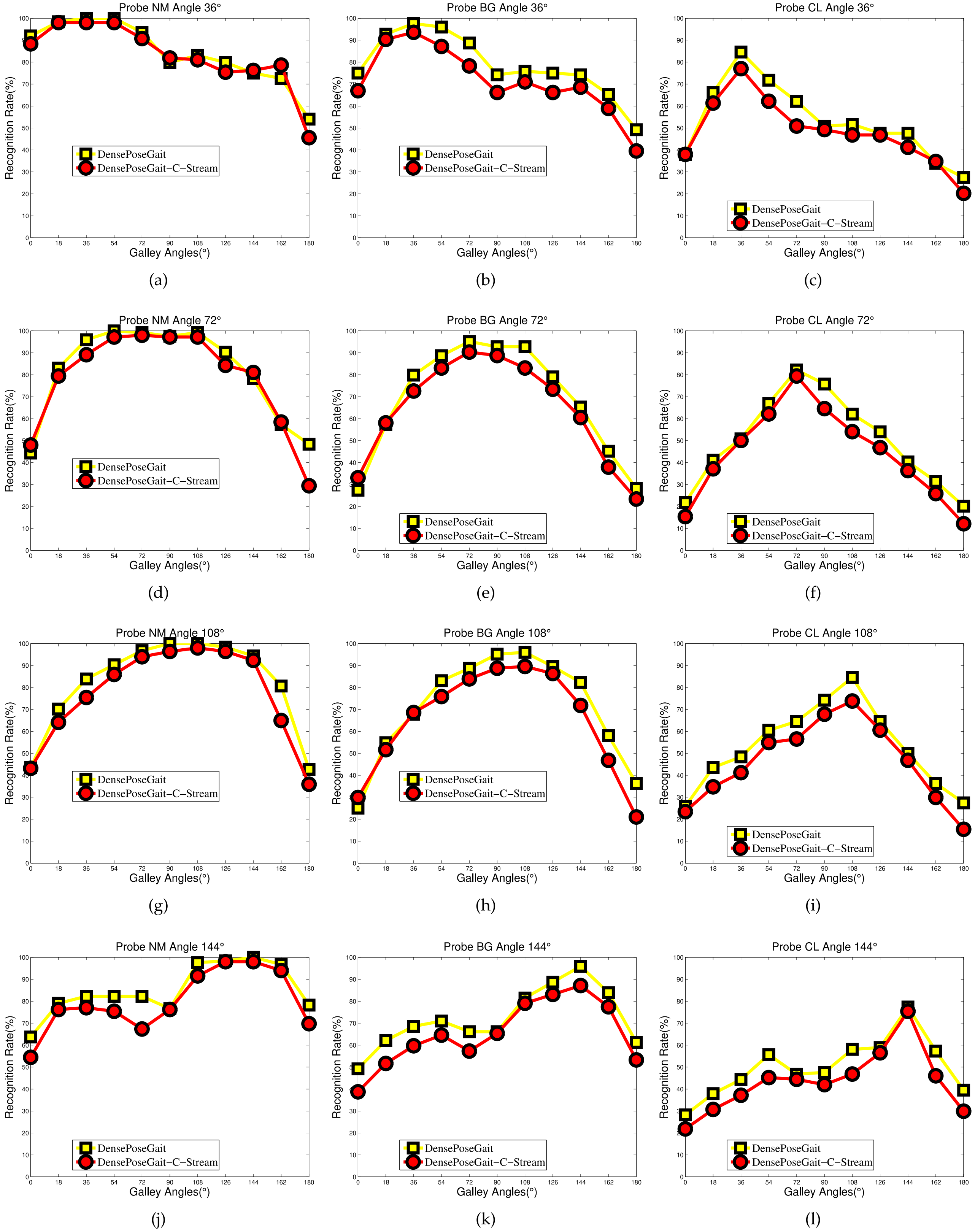


Fig. 10. Comparing with cross-view methods at probe angle 54°, 90° and 126°. Galley angles are from 0° to 180°.

Looking forward, leveraging the framework of *DensePoseGait*, with the refinement of network, pose estimation techniques, and the enrichment of gait datasets, it



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 11. Experimental results of *DensePoseGait-C-Stream* and *DensePoseGait*. From left column to right column are NM, BG and CL condition respectively.

is anticipated that *DensePoseGait*'s performance will continue to ascend, fortifying its position as a transformative approach in model-based gait recognition methodologies.

D. Effectiveness on View Variation

The input feature of the proposed method *DensePoseGait* consists of rich UV coordinates, and UV coordinates contain

3D body informants. In order to analyze the effectiveness of the view variation, we compared our method *DensePoseGait* with some cross-view gait recognition methods. Including FD-VTM [41], RSVD-VTM [42], RPCA-VTM [43], R-VTM [44], GP+CCA [45] and C3A [46]. Those methods use view transformation model (VTM) to reduce the effect of view variation. VTM can transform gait template features from one view to another view for improving the robustness of the view. Three probe angles (54°, 90°, and 126°) are chosen to compare. The recognition rates are shown in Fig. 10.

From Fig. 10, it is obvious that our proposed method, *DensePoseGait*, demonstrates exceptional performance when there is a significant angle difference between the gallery and probe. As the angle difference increases, so does the improvement in performance. This highlights the advantage of *DensePoseGait*, as it is specifically designed for modeling 3D human body movement, making it more robust to variations in viewpoint.

E. Ablation Study

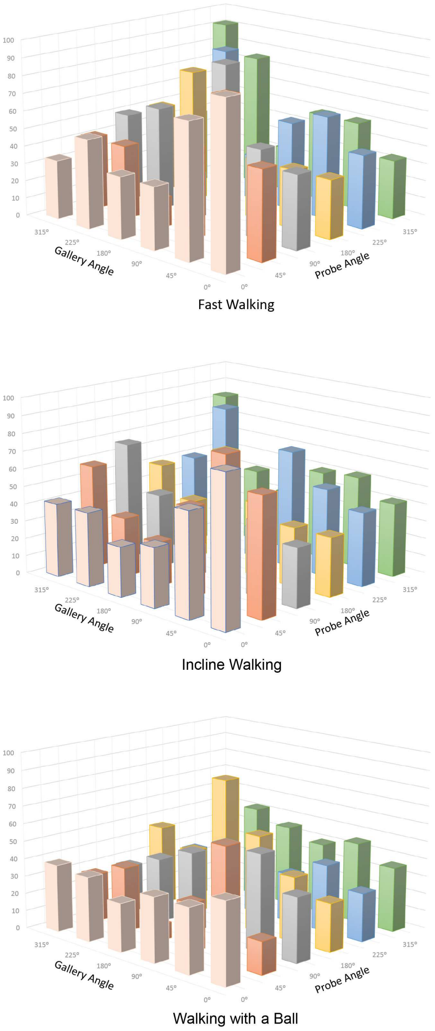
In order to show the proposed framework can further promote the learning of gait features. We only use the C-Stream to train a model, namely *DensePoseGait-C-Stream*, while the *DensePoseGait* is trained by using C-Stream and with the guide learning with TP-Stream and LP-Stream. For the reference stage, both are using C-Stream. The experimental results can be shown in Fig. 11. Due to limited space, we only list 4 probe angles with a 36° interval, that is, 36°, 72°, 108° and 144°. From Fig. 11, we can see that the performance of *DensePoseGait* is better than that of *DensePoseGait-C-Stream* at many points. This is because during the training stage, the representative features would become better with the guidance of TP-Stream and LP-Stream. It shows that the human body parts have a positive influence on the global feature learning.

F. Experimental Results on MoBo Dataset

The MoBo dataset experimental results can be found in Fig. 12. This figure displays the evaluation results under different scenarios such as varying views, fast walking, incline walking, and walking with a ball. In the experiment, slow walking sequences were placed in the gallery set while fast walking, incline walking, and walking with a ball were placed in the probe set. Each set of experiments contains 36 combinations, resulting in 36 recognition rates per figure.

G. Comparisons on MoBo Dataset

Except for experiments on CASIA-B dataset, we also perform evaluation experiments and comparisons on the MoBo dataset. The comparison methods consists of model-based methods PoseGait [9] and PoseMapGait [13], and appearance-based methods GaitGANv2 [25], DV-GEIs-pre [4] and DV-GEIs [3]. We conducted above methods by ourselves because they do not perform the experiments on the MoBo dataset according to the original paper. To gain a better overview of the comparisons, we segregated them into two categories depending on the variations in their conditions.



Fast Walking

Incline Walking

Walking with a Ball

Fig. 12. The experimental results under three conditions on MoBo dataset.

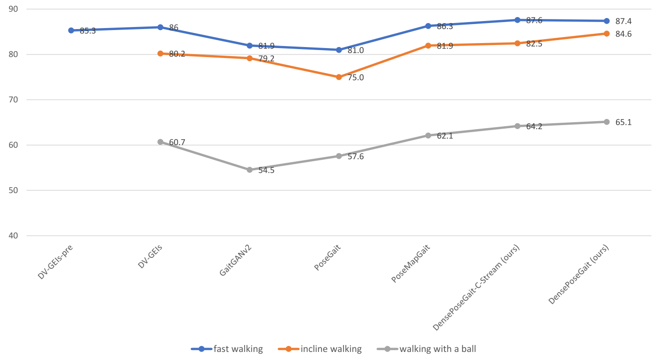


Fig. 13. The average recognition rates for the probe data being fast walking, incline walking and walking with a ball on identical-view cases.

Identical-View Comparison: The mean recognition rates on identical-view cases can be shown in Fig. 13. From Fig. 13, it is obvious that proposed methods can achieve better accuracy than model-based methods and appearance-based methods, which shows that dense pose features has a big potential to improve the robustness of gait recognition on the real world.

Cross-View Comparison: The mean recognition rates on the cross-view cases can be shown in Table VI. We can see that our proposed methods not only can perform better on

TABLE VI
AVERAGE ACCURACIES (%) ON MOBO DATASET UNDER THREE DIFFERENT EXPERIMENTAL SETTINGS, EXCLUDING IDENTICAL-VIEW CASES

Gallery angle (slow walking)	0°, 45°, 90°, 180°, 225°, 315°						
Probe angle (fast walking)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [25]	31.7	46.7	50.0	33.3	36.7	50.0	41.4
DV-GEIs [3]	41.7	48.3	48.3	38.3	51.7	53.3	46.9
PoseGait [9]	38.3	45.0	43.3	25.0	36.7	45.0	38.9
PoseMapGait [13]	48.3	43.3	43.3	40.0	46.7	58.3	46.7
DesenPoseGait-C-Stream (ours)	49.8	45.8	47.6	38.5	47.7	62.3	48.6
DesenPoseGait (ours)	50.1	51.2	47.6	40.0	46.8	57.1	**48.8**
Probe angle (incline walking)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [25]	36.7	51.7	38.3	31.7	43.3	46.7	41.4
DV-GEIs [3]	40.0	51.7	36.7	35.0	38.3	51.7	42.2
PoseGait [9]	36.7	50.0	36.7	30.0	35.0	51.7	40.0
PoseMapGait [13]	40.0	46.7	40.0	38.3	45.0	45.0	42.5
DesenPoseGait-C-Stream (ours)	46.0	45.3	39.0	39.6	45.6	47.3	43.8
DesenPoseGait (ours)	36.5	48.6	41.3	48.3	43.6	51.2	**44.9**
Probe angle (walking with a ball)	0°	45°	90°	180°	225°	315°	Mean
GaitGANv2 [25]	32.7	27.3	30.9	27.3	32.7	27.3	29.7
DV-GEIs [3]	38.2	27.3	32.7	41.8	40.0	36.4	36.1
PoseGait [9]	32.7	21.8	32.7	23.6	38.2	34.5	30.6
PoseMapGait [13]	36.4	30.9	32.7	45.5	41.8	40.0	37.9
DesenPoseGait-C-Stream (ours)	34.9	32.6	39.0	39.6	45.6	43.6	39.2
DesenPoseGait (ours)	36.5	35.6	40.3	48.3	43.6	42.8	**41.2**

the identical-view cases, but also on the cross-view cases. In addition, the comparison of *DesenPoseGait-C-Stream* and *DesenPoseGait* also can further show that the human body parts has a positive impact on the feature learning of global.

V. CONCLUSION AND FUTURE WORK

Our study introduced the *DensePoseGait* framework, a novel approach in model-based gait recognition that leverages dense pose maps for initial input data. This approach marks a significant advancement over traditional skeleton-based and silhouette-based methods. The dense pose maps provide a richer and more robust 3D pose representation, particularly effective against variations in human shape. Our experiments on the CASIA-B and MoBo datasets have demonstrated that *DensePoseGait* sets a new benchmark in state-of-the-art performance for model-based gait recognition systems.

A key insight from our research is the pivotal role of initial data in gait recognition systems. The success of *DensePoseGait* highlights how advanced data representations, such as dense pose maps, can substantially improve robustness and accuracy, particularly in scenarios with significant gait variation. This finding can be instrumental in guiding future research towards exploring and developing more sophisticated data types for gait recognition.

Looking forward, we identify several promising directions for further research:

Exploration of 3D Point Cloud Data: Building on the success of dense pose maps, we propose investigating the use of 3D point cloud data in gait recognition. As camera technology and pose estimation algorithms continue to evolve, 3D point cloud data could offer even more detailed and accurate representations of human gait, such as LidarGait [47], potentially opening new avenues for research and application.

Enhancement of Pose Estimation Algorithms: Continuous improvements in pose estimation algorithms will be crucial for the advancement of gait recognition technology. We plan to contribute to this area by developing more sophisticated algorithms that can accurately capture subtle gait nuances, further enhancing the performance of systems like *DensePoseGait*. Additionally, we are considering integrating complementary technologies, such as depth sensors or advanced texture mapping, to enhance the detail captured in our 3D body surface model. This integration could potentially address the limitations you highlighted and further improve the robustness and accuracy of our gait recognition approach.

Application in Real-world Scenarios: Given the non-contact and long-distance identification advantages of gait recognition, particularly highlighted during the COVID-19 pandemic, we aim to test and refine DensePoseGait in various real-world scenarios. This includes deployment in surveillance, healthcare monitoring, and human-computer interaction, to assess its practicality and efficacy in dynamic environments.

In conclusion, DensePoseGait represents a significant step forward in gait recognition technology. With ongoing research and development, we anticipate making substantial contributions to the field, enhancing the utility of gait recognition in various applications, and addressing emerging challenges in a world increasingly reliant on sophisticated biometric technologies.

ACKNOWLEDGMENT

The views expressed in this article, book, or presentation are those of the author and do not necessarily reflect the official policy or position of the United States Air Force Academy, the Air Force, the Department of Defense, or the U.S. Government. Approved for public release: distribution unlimited. PA Number USAFA-DF-2023-79.

REFERENCES

[1] X. Yang, Z. Wang, H. Wu, L. Jiao, Y. Xu, and H. Chen, "Stable and compact face recognition via unlabeled data driven sparse representation-based classification," *Signal Process., Image Commun.*, vol. 111, Feb. 2023, Art. no. 116889.

[2] J. Tao and Y.-P. Tan, "A probabilistic approach to incorporating domain knowledge for closed-room people monitoring," *Signal Process., Image Commun.*, vol. 19, no. 10, pp. 959–974, 2004.

[3] R. Liao, W. An, Z. Li, and S. S. Bhattacharyya, "A novel view synthesis approach based on view space covering for gait recognition," *Neurocomputing*, vol. 453, pp. 13–25, Sep. 2021.

[4] R. Liao, W. An, S. Yu, Z. Li, and Y. Huang, "Dense-view GEIs set: View space covering for gait recognition based on dense-view GAN," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, 2020, pp. 1–9.

[5] R. Liao, Z. Li, S. S. Bhattacharyya, and G. York, "View DiffGait: View pyramid diffusion for gait recognition," in *Proc. IEEE 18th Int. Conf. Automat. Face Gesture Recognit. (FG)*, 2024, pp. 1–9.

[6] 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," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 474–483.

[7] W. An, R. Liao, S. Yu, Y. Huang, and P. C. Yuen, "Improving gait recognition with 3D pose estimation," in *Proc. Chin. Conf. Biometric Recognit.*, 2018, pp. 137–147.

[8] W. An et al., "Performance evaluation of model-based gait on multi-view very large population database with pose sequences," *IEEE Trans. Biometrics, Behav., Ident. Sci.*, vol. 2, no. 4, pp. 421–430, Oct. 2020.

[9] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069.

[10] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7297–7306.

[11] C. Fan et al., "Gaitpart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14213–14221.

[12] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognit.*, vol. 43, no. 6, pp. 2281–2291, 2010.

[13] 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*, vol. 501, pp. 514–528, Aug. 2022.

[14] C. Liu, R. Zhang, and L. Guo, "Part-pose guided Amur tiger re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019, pp. 315–322.

[15] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2001, pp. II–II.

[16] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. ICCV*, 2017, pp. 2353–2362.

[17] L. Wang, H. Ning, T. Tan, and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 2, pp. 149–158, Feb. 2004.

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

[19] P. Schwarz, J. Scharinger, and P. Hofer, "Gait recognition with densePose energy images," in *Proc. Int. Conf. Syst., Signals Image Process.*, 2021, pp. 65–70.

[20] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu, "SkeletonGait: Gait recognition using skeleton maps," 2023, *arXiv:2311.13444*.

[21] Z. Wang, S. Hou, M. Zhang, X. Liu, C. Cao, and Y. Huang, "GaitParsing: Human semantic parsing for gait recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 4736–4748, 2024.

[22] J. Zheng, X. Liu, S. Wang, L. Wang, C. Yan, and W. Liu, "Parsing is all you need for accurate gait recognition in the wild," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 116–124.

[23] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[24] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2164–2176, Nov. 2012.

[25] S. Yu et al., "GaitGANv2: Invariant gait feature extraction using generative adversarial networks," *Pattern Recognit.*, vol. 87, pp. 179–189, Mar. 2019.

[26] J. Li, J. Gao, Y. Zhang, H. Shan, and J. Zhang, "Motion matters: A novel motion modeling for cross-view gait feature learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2023, pp. 1–5.

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

[28] J. Chen, Z. Wang, P. Yi, K. Zeng, Z. He, and Q. Zou, "Gait pyramid attention network: Toward silhouette semantic relation learning for gait recognition," *IEEE Trans. Biometrics, Behav., Ident. Sci.*, vol. 4, no. 4, pp. 582–595, Oct. 2022.

[29] J. Chen, Z. Wang, C. Zheng, K. Zeng, Q. Zou, and L. Cui, "GaitAMR: Cross-view gait recognition via aggregated multi-feature representation," *Inf. Sci.*, vol. 636, Jul. 2023, Art. no. 118920.

[30] H. Dou, P. Zhang, Y. Zhao, L. Dong, Z. Qin, and X. Li, "GaitMPL: Gait recognition with memory-augmented progressive learning," *IEEE Trans. Image Process.*, vol. 33, pp. 1464–1475, 2024.

[31] 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 Trans. Comput. Vis. Appl.*, vol. 10, no. 4, pp. 1–14, 2018.

[32] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit.*, 2006, pp. 441–444.

[33] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Carnegie Mellon Univ., Pittsburgh, PA, USA, Rep. CMU-RI-TR-01-18, 2001. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7506114

[34] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, pp. 1511–1521, 2012.

[35] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 20196–20205.

[36] Z. Zhu et al., "Gait recognition in the wild: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14789–14799.

[37] S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, "Invariant feature extraction for gait recognition using only one uniform model," *Neurocomputing*, vol. 239, pp. 81–93, May 2017.

[38] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 532–539.

[39] Z. Huang et al., "3D local convolutional neural networks for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 14920–14929.

[40] X. Huang et al., "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12909–12918.

[41] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 151–163.

[42] W. Kusakunniran, Q. Wu, H. Li, and J. Zhang, "Multiple views gait recognition using view transformation model based on optimized gait energy image," in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops*, 2009, pp. 1058–1064.

[43] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *Proc. 18th IEEE Int. Conf. Image Process.*, 2011, pp. 2073–2076.

[44] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012.

[45] K. Bashir, T. Xiang, and S. Gong, "Cross view gait recognition using correlation strength," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–11.

[46] X. Xing, K. Wang, T. Yan, and Z. Lv, "Complete canonical correlation analysis with application to multi-view gait recognition," *Pattern Recognit.*, vol. 50, pp. 107–117, Feb. 2016.

[47] C. Shen, C. Fan, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "LidarGait: Benchmarking 3D gait recognition with point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1054–1063.

Rijun Liao received the B.S. and M.S. degrees from the College of Computer Science and Software Engineering, Shenzhen University, China, in 2015 and 2018, respectively, and the Ph.D. degree from the Department of Computer Science and Electrical Engineering, University of Missouri–Kansas City in 2024. He is currently an Assistant Professor with the School of Biomedical Engineering, Guangdong Medical University. His research area includes machine learning, computer vision, image processing, and natural language processing. Based on these researches, he published more than 10 publications and received more than 1000 citations in his google scholar. He also serves as a Reviewer for IEEE Transactions on Image Processing, IEEE Transactions on Human-Machine Systems, and IEEE International Conference on Acoustics, Speech and Signal Processing.

Zhu Li (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Northwestern University, Evanston, in 2004. He is currently a Professor with the Department of Computer Science and Electrical Engineering, University of Missouri, Kansas City, and the Director of the NSF I/UCRC Center for Big Learning, UMKC. He was an AFOSR SFFP Summer Visiting Faculty with the UAV Research Center, U.S. Air Force Academy in 2016, 2017, 2018, and 2020. He was a Sr. Staff Researcher/Sr. Manager with Samsung Research America's Multimedia Standards Research Lab, Richardson, TX, USA, from 2012 to 2015, a Sr. Staff Researcher/Media Analytics Lead with FutureWei (Huawei) Technology's Media Lab, Bridgewater, NJ, USA, from 2010 to 2012, an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University from 2008 to 2010, and a Principal Staff Research Engineer with the Multimedia Research Lab, Motorola Labs from 2000 to 2008. His research interests include point cloud and light field compression, graph signal processing, and deep learning in the next gen visual compression, image processing, and understanding. He has 47 issued or pending patents, 100+ publications in book chapters, journals, and conferences in these areas. He received the Best Paper Award at IEEE International Conference on Multimedia and Expo, Toronto, in 2006, and the Best Paper Award (DoCoMo Labs Innovative Paper) at IEEE International Conference on Image Processing, San Antonio, in 2007. He is an Associate Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING in 2020, IEEE TRANSACTIONS ON MULTIMEDIA from 2015 to 2018, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2016 to 2019. He serves on the Steering Committee Member of IEEE ICME from 2015 to 2018. He is an Elected Member of the IEEE Multimedia Signal Processing, IEEE Image, Video, and Multidimensional Signal Processing, and IEEE Visual Signal Processing and Communication Tech Committees. He is the Program Co-Chair for IEEE International Conference on Multimedia and Expo in 2019, and co-chaired the IEEE Visual Communication and Image Processing in 2017.

Shuvra S. Bhattacharyya (Fellow, IEEE) received the Ph.D. degree from the University of California at Berkeley. He is a Professor with the Department of Electrical and Computer Engineering, The University of Maryland at College Park, College Park. He holds a joint appointment with the University of Maryland Institute for Advanced Computer Studies. He also holds a part-time position as the International Research Chair joint with INSA/IETR, and INRIA, Rennes, France. He has held industrial positions as a Researcher with the Hitachi America Semiconductor Research Laboratory, San Jose, CA, USA, and a Compiler Developer with Kuck and Associates, Champaign, IL, USA. He has held a visiting summer research position with AFRL, Rome, NY, USA. From 2015 to 2018, he was a part-time Visiting Professor with the Department of Pervasive Computing, Tampere University of Technology, Finland, as part of the Finland Distinguished Professor Programme. His research interests include signal processing, embedded systems, electronic design automation, machine learning, wireless communication, and wireless sensor networks.

George York (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Washington. He is a Professor of Electrical and Computer Engineering with the United States Air Force Academy, CO, USA, and is the Director of the Academy Center for Unmanned Aircraft Systems Research. He has 50+ publications in book chapters, journals, and conference proceedings. His research interests include the cooperative control of multiple autonomous systems, digital signal processing, and embedded computer systems. He serves on the National Council of Examiners for Engineering and Surveying Electrical and Computer Professional Engineering exam subcommittee.