applied sciences

MDPI

Article

Two-Path Spatial-Temporal Feature Fusion and View Embedding for Gait Recognition

Diyuan Guan [1,2], Chunsheng Hua [1,*] and Xiaoheng Zhao [1]

[1] Institute of Intelligent Robots and Pattern Recognition, College of Information, Liaoning University, Shenyang 110036, China
[2] College of Information Engineering, Shenyang University, Shenyang 110044, China
* Correspondence: huachunsheng@lnu.edu.cn

Abstract: Gait recognition is a distinctive biometric technique that can identify pedestrians by their walking patterns from considerable distances. A critical challenge in gait recognition lies in effectively acquiring discriminative spatial-temporal representations from silhouettes that exhibit invariance to disturbances. In this paper, we present a novel gait recognition network by aggregating features in the spatial-temporal and view domains, which consists of two-path spatial-temporal feature fusion module and view embedding module. Specifically, two-path spatial-temporal feature fusion module firstly utilizes multi-scale feature extraction (MSFE) to enrich the input features with multiple convolution kernels of various sizes. Then, frame-level spatial feature extraction (FLSFE) and multi-scale temporal feature extraction (MSTFE) are parallelly constructed to capture spatial and temporal gait features of different granularities and these features are fused together to obtain muti-scale spatial-temporal features. FLSFE is designed to extract both global and local gait features by employing a specially designed residual operation. Simultaneously, MSTFE is applied to adaptively interact multi-scale temporal features and produce suitable motion representations in temporal domain. Taking into account the view information, we introduce a view embedding module to reduce the impact of differing viewpoints. Through the extensive experimentation over CASIA-B and OU-MVLP datasets, the proposed method has achieved superior performance to the other state-of-the-art gait recognition approaches.

Keywords: gait recognition; two-path spatial-temporal feature fusion; multi-scale feature extraction; view embedding

Citation: Guan, D.; Hua, C.; Zhao, X. Two-Path Spatial-Temporal Feature Fusion and View Embedding for Gait Recognition. *Appl. Sci.* **2023**, *13*, 12808. https://doi.org/10.3390/app132312808

Academic Editors: Md. Shohel Sayeed, Tee Connie and Ong Thian Song

Received: 12 November 2023
Revised: 28 November 2023
Accepted: 28 November 2023
Published: 29 November 2023

1. Introduction

Gait recognition refers to judging individuals by their distinct walking patterns, making it one of the most auspicious biometric technologies for identity recognition. In contrast to other biometric identification methods such as fingerprints, DNA, facial recognition, vein recognition, and so on, gait recognition works by using regular or low-resolution cameras at long ranges and does not require explicit cooperation from the subjects of interest. Therefore, gait recognition has demonstrated huge development potential in the fields of crime prevention, video surveillance, and social security. Invariance to disturbances in gait recognition signifies the capability of achieving high recognition accuracy without being affected by various external factors, such as bag-carrying, cloth-change, cross-view, and speed changes [1–3]. However, the performance of gait recognition is susceptible to the forementioned external factors in real-word scenarios, which brings substantial challenges to achieve invariance to disturbances. Therefore, a multitude of state-of-art works have focused on achieving invariance to disturbances in gait recognition to ensure the reliability of gait recognition systems.

Benefiting from recent developments in deep learning algorithms, numerous powerful gait recognition algorithms have been developed [4–11], and have been proven

effective under challenging conditions. Liao et al. [9] proposed long short-term memory (LSTM) [12] for the extraction of spatial and temporal features based on human skeleton points. PoseGait [10] transformed 2D poses into 3D poses to enhance recognition accuracy by extracting additional gait features from 3D poses. Shiraga et al. [11] utilized the gait energy image (GEI) to train their networks and learn covariate features for gait recognition. GaitSet [4] used 2D convolutional neural networks (CNNs) at the frame level to extract global features and treated gait silhouettes as a set to capture temporal features. GaitPart [7] extracted local gait feature representations by dividing the feature maps horizontally and utilized micro-motion features to focus on the short-term temporal expressions. Huang et al. [8] introduced a 3D local CNN to extract sequence information from specific human body parts.

To the best of our knowledge, the human body possess evidently various visual appearances and movement patterns during walking. Spatial-temporal representations refer to the spatial-temporal features derived from the modeling of silhouettes, which can effectively capture the visual appearances and movement patterns of the pedestrian. Spatial feature representations indicate the visual appearances of the silhouettes, while temporal feature representations reflect the movement patterns of the silhouettes. Spatial feature extraction and temporal modeling can extract rich and discriminative spatial-temporal feature representations, distinctly capturing an individual's walking process. Since there are significant differences among various persons in terms of visual appearance and movement patterns during walking, these spatial-temporal representations can play a pivotal role in enhancing the effectiveness of gait recognition. The combing of spatial and temporal representations can not only describe the visual appearances, but also capture motion patterns of the pedestrians, and utilizing either spatial representations or temporal representations alone would lead to the poor accuracy of gait recognition. Therefore, by jointly investigating motion learning and spatial mining simultaneously, the accuracy of gait recognition can be substantially improved. Despite considerable efforts in gait recognition, the aforementioned methods still encounter the following challenges: (1) there is a need for muti-scale feature extraction to capture more robust spatial-temporal representations, thereby enhancing the accuracy of gait recognition especially in case of appearance camouflage; (2) few methods have taken view angle into consideration explicitly and the detection or estimation of viewpoint has been somewhat overlooked, which can exert an essential impact to improve the recognition ability of existing approaches.

With such considerations, we propose an advanced gait recognition network, namely two-path spatial-temporal feature fusion module and view embedding module. The two-path spatial-temporal feature fusion module consists of multi-scale feature extraction (MSFE), frame-level spatial feature extraction (FLSFE) and multi-scale temporal feature extraction (MSTFE). Firstly, MSFE is deployed to facilitate the effective extraction of shallow features, which can extend the receptive field and enable the extraction of multiple internal features within different regions. Subsequently, we introduce a two-path parallel structure containing FLSFE and MSTFE, which aims to effectively extract the multi-scale spatial-temporal information across various granularities. In FLSFE, we develop an innovative residual convolutional block (R-conv) to capture both global and local gait features by the special design of residual operation. Meanwhile, in MSTFE, we design independent branches of temporal feature extraction with varying scales and integrate the temporal features in an attention-based way. For reasonable refinement of extracted features, a view embedding module is constructed to reduce the negative impact of viewpoint differences, which uses view prediction learning to calculate the best view and embeds the view information into the multi-scale spatial-temporal characteristics to obtain the final features. Extensive experiments conducted on the two public datasets CASIA-B and OU-MVLP have demonstrated that our method has outperformed other state-of-the-art methods, showing its superior performance in gait recognition.

2. Related Work

Gait recognition: Current deep-learning-based gait recognition methods can be broadly classified into two categories: model-based [13–16] and appearance-based [4–8,17–30]. Model-based methods leverage the relationships between bone joints and pose information to create models of walking patterns and human body structures [13], such as OpenPose [14], HR-Net [15], and DensePose [16]. These methods exhibit stronger robustness to clothing variation and carrying articles. However, model-based methods depend on accurate joint detection and pose estimators, which can significantly increase the computational complexity and may lead to inferior performance in certain scenarios. Conversely, appearance-based methods use gait silhouettes (the binary images shown in Figure 1) as the model's input and capture spatial and temporal characteristics from silhouettes by CNNs [17]. Gait silhouettes can describe the body state in a single frame at a low computational cost and more detailed information in each silhouette image can be preserved directly from the original silhouette sequences. The silhouettes are the basis of appearance-based methods, and the quality of silhouettes will directly affect the performance of the gait recognition system. Moreover, spatial feature representations can be obtained from the silhouette of an individual frame, which can represent appearance characteristics, while temporal feature representations can be captured from consecutive silhouettes, in which the relationship between adjacent frames can reflect the temporal characteristics and motion patterns. Therefore, some appearance-based methods have overcome the challenges of pose estimation and achieved competitive performance [4–8,21,22,29]. Particularly, the first opensource gait recognition framework, named OpenGait (https://github.com/ShiqiYu/OpenGait, accessed on 29 January 2023) [18], encompassed a series of state-of-the-art appearance-based methods for gait recognition. In this paper, our approach is categorized as appearance-based and we count on binary silhouettes as our input data without the need for pose estimation or joint detection. By focusing on silhouettes, we aim to reduce the influence of variations in subjects' appearance, thereby enhancing the accuracy and robustness of our gait recognition approach.

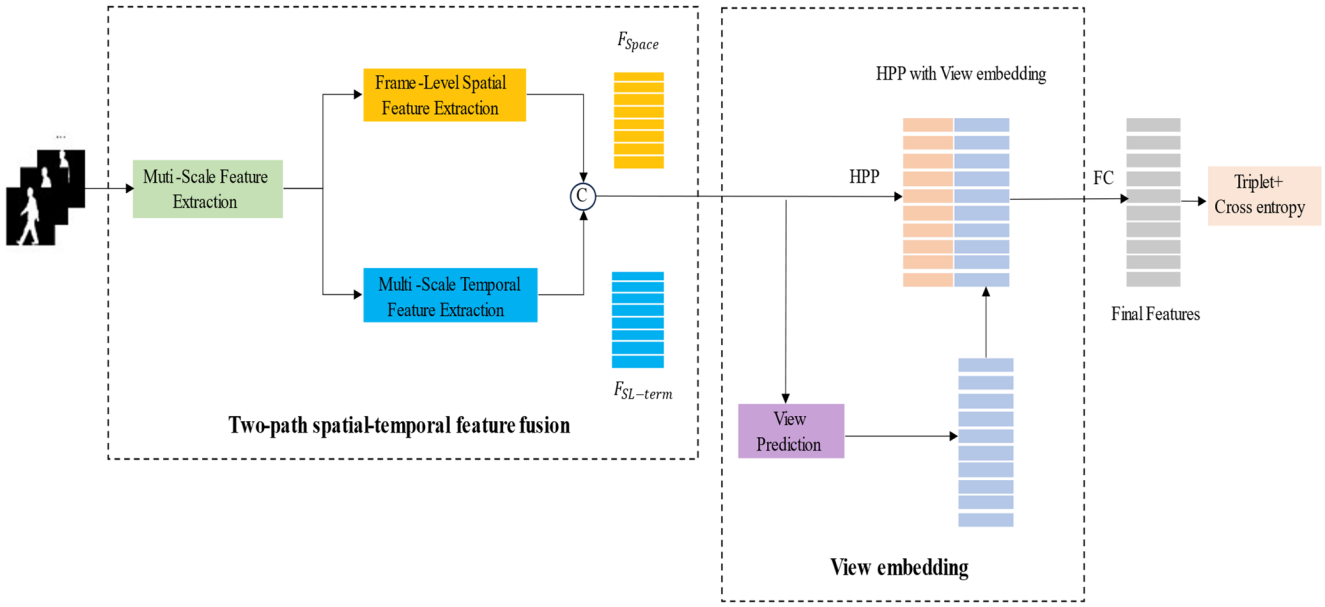

Figure 1. Overview of the whole framework, which mainly consists of two modules as: two-path spatial-temporal feature fusion module and view embedding module. The two-path spatial-temporal feature fusion module consists of three component including multi-scale feature extraction (MSFE), frame-level spatial feature extraction (FLSFE) and multi-scale temporal feature extraction (MSTFE). View embedding module consists of two components: view prediction and HPP with view embedding.

Spatial feature extraction modeling: Regarding the range of feature representations in spatial feature extraction modeling, two main approaches are commonly used: global-based and local-based methods. Specifically, global-based methods focus on exploring gait silhouettes as a whole to generate global feature representations [4,5,12,17,19,20]. For instance, Shiraga et al. [12] proposed the GEI and utilized 2D CNNs to obtain global gait feature representations from the GEIs. Similarly, GaitSet [4] and GLN [5] extracted global gait features at the frame-level by 2D CNNs. Conversely, local-based methods usually segment the silhouettes into multiple parts to establish the local feature representations [7,21,22], and focus more on learning the local information of different body parts. For example, Zhang et al. [21] split human gait into four distinct segments and adopted 2D CNNs to capture more detailed information from each of these segments. Fan et al. [7] introduced the focal convolution layer (FConv), a novel convolution layer that divided the feature map into several parts to obtain part-based gait features. Qin et al. [22] developed RPNet to discover the intricate interconnections between each part of gait silhouettes and then integrated them by a straightforward splicing process. GaitGL [23] and GaitStrip [24] both used 3D CNNs to construct multi-level frameworks to extract more discriminative and richer spatial features. In this paper, we design an innovative residual convolutional block (R-conv) in spatial feature extraction model by 2D CNNs, which combines regular convolution with FConv to extract both global and local gait features and enhance the discriminative capacity of the feature representations.

Temporal feature extraction modeling: As a crucial cue for gait tasks, the current temporal feature extraction model generally employs various approaches such as 1D convolutions, LSTMs and 3D convolutions [25]. For example, Fan et al. [7] and Wu et al. [26] utilized 1D convolutions to model the temporal dependencies and aggregated temporal information by concatenation or in a summation. Additionally, LSTM networks were built in [21,27] to preserve the temporal variation of gait sequences and fuse temporal information by accumulation. Moreover, some studies have proposed 3D convolutions [28–31] to simultaneously extract spatial and temporal information from gait silhouette sequences. Lin et al. [32] introduced MT3D network, which used 3D CNNs to extract spatial-temporal features with multiple temporal scales. However, 3D CNNs often bring complex calculations and encounter challenges during training. In this paper, we present a novel approach for temporal feature extraction modeling by 2D CNNs, which aggregates temporal information with different scales. By incorporating muti-scale temporal branches, our approach can capture rich temporal clues and empower the network to learn more discriminative motion representations adaptively.

View-invariant Modeling: To the best of our knowledge, viewpoint change poses a formidable challenge in biometrics, particularly in face recognition and gait recognition. In contrast to face recognition, fewer methods in gait recognition have incorporated the aspect of viewpoint into their considerations. He et al. [33] introduced a multitask generative adversarial network (GAN) and trained the GAN by using viewpoint labels as the supervision. Chai et al. [34] adopted a different projection matrix as a perspective embedding method and achieved high growth on multiple backbones. However, these methods often involve a large number of parameters, making them extremely complex for effective cross-view gait recognition. Therefore, we propose a concise view model that applies view prediction learning to calculate the best view and embeds view information into our two-path spatial-temporal feature fusion module, which can significantly enhance the robustness of our network to view changes and improve gait recognition performance across varying viewpoints.

3. Proposed Method

3.1. System Overview

The overall framework is depicted in Figure 1. Firstly, MSFE is employed to extract shallow features from the original input silhouettes, which can extend the receptive field and capture the gait spatial and temporal features within different regions. Next, a two-path

parallel structure including FLSFE and MSTFE is designed to obtain multi-scale spatial-temporal features. FLSFE is devised to extract combined features by encompassing both global and local gait information in the spatial domain and use a novel convolutional block (R-conv) by the special design of residual operation. Simultaneously, MSTFE is implemented to learn more discriminative motion representation between long-term and short-term features in the temporal domain and integrate the muti-scale temporal features in an attention-based way. Subsequently, both the spatial and temporal features are integrated to generate muti-scale spatial-temporal representations. HPP [4] is applied to complete feature mapping process, ensuring comprehensive and discriminative features. Furthermore, a view embedding module is introduced to use view prediction learning to calculate the best view and incorporate view information explicitly into the spatial-temporal fused feature representations to gain the final features, which can effectively alleviate the effect of viewpoint variations. Finally, joint losses of cross entropy loss and the triplet loss [4,7] are selected to train the proposed network.

3.2. Two-Path Spatial-Temporal Feature Fusion Module

The two-path spatial-temporal feature fusion module is composed of three different components, namely multi-scale feature extraction (MSFE), frame-level spatial feature extraction (FLSFE) and multi-scale temporal feature extraction (MSTFE). In the following sections, we will furnish a comprehensive description of the structure of each component individually.

3.2.1. Multi-Scale Feature Extraction (MSFE)

Generally, a convolution layer utilizes multiple convolution kernels of the same size to carry out convolution operation, which leads to the limitation of feature extraction to a certain extent. In the two-path spatial-temporal feature fusion module, we propose multi-scale feature extraction (MSFE) to process the input silhouettes, which can extract more comprehensive and discriminative feature representations with different granularities. MSFE utilizes the multi-scale convolution (MSC) structure with multiple kernels of various sizes to extract gait features during convolution operation and merges the multi-scale features by different single-path convolution branches. The MSC can expand different perceptual fields and learn more fine-grained features. Moreover, multiple branches can fuse the fine-grained features and obtain more abundant and complete gait characteristics.

The structure of MSFE is presented in Figure 2. MSFE comprises a three-layer structure, where the first layer is a MSC structure and the other two layers are the regular convolutions. The MSC structure comprises three parallel convolution branches, each of which is designed with distinct kernel sizes: (1,1), (3,3), and (5,5), respectively. Each branch processes the input gait sequences and transforms them into multi-scale feature maps. Then, the outputs of the parallel convolution operations at different scales are combined together by element-wise addition to obtain more abundant features. Subsequently, another two convolution layers are employed to optimize the comprehensive characteristics to achieve the appropriate number of channels, which can also balance the capabilities of the network more flexibly. Assuming that the input is $X \in R^{C \times T \times H \times W}$, where C means the number of channels, T represents the length of the gait sequence and (H, W) signify the height and width of each frame. The MSC feature F_{MSC} can be formulated below:

$$F_{MSC} = F_{1\times1}(X) + F_{3\times3}(X) + F_{5\times5}(X) \tag{1}$$

where $F_{1\times1}$, $F_{3\times3}$ and $F_{5\times5}$ denote the 2D convolution operations with the kernel sizes of (1,1), (3,3) and (5,5), respectively.

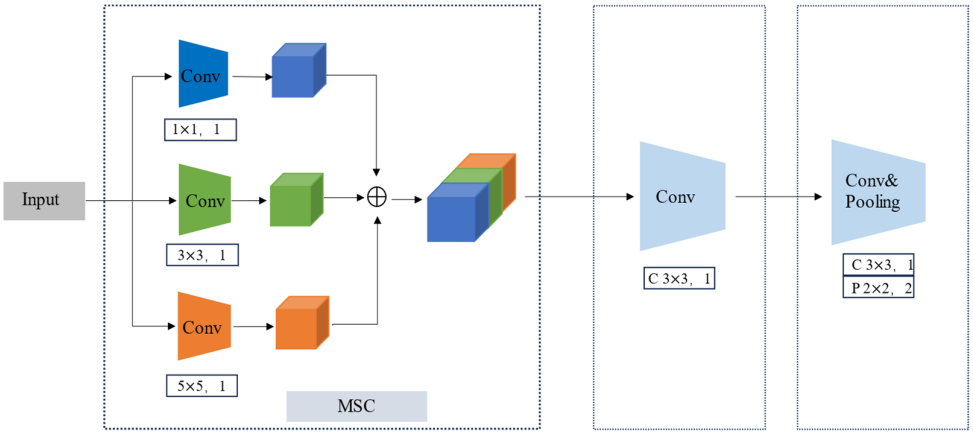

Figure 2. The detailed structure of multi-scale feature extraction (MSFE).

3.2.2. Frame-Level Spatial Feature Extraction (FLSFE)

To further enhance the granularity of spatial-temporal learning and capture contextual information, frame-level spatial feature extraction (FLSFE) is constructed to extract both global and local spatial features for each frame. The structure of FLSFE is shown in Figure 3. The input of FLSFE is demoted as $F_{MSFE} \in R^{C\times T\times H/2\times w/2}$, and FLSFE is composed of three consecutive R-convs (detailed structure could be found in Figure 4), each of which is designed to extract both the whole-body and part-informed spatial features. Then, set pooling operation (SP) is applied to each R-conv block to extract set-level features, and after point-by-point summation, column vector spatial features $F_{space} \in R^{C_1\times K}$ are obtained by the horizontal pooling operation (HP), where C_1 means the number of channels, and K represents the number of strips cut in HP.

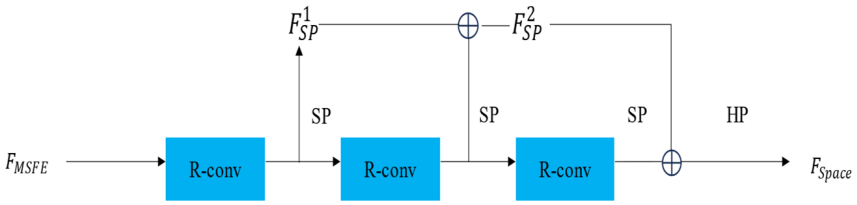

Figure 3. The detailed structure of frame-level spatial feature extraction (FLSFE).

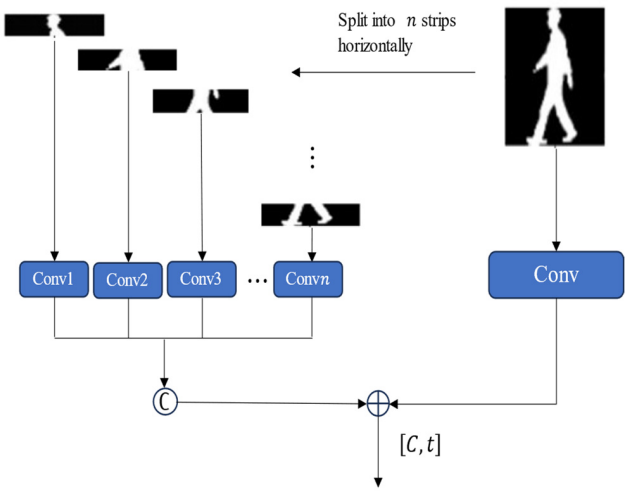

Figure 4. The detailed structure of R-conv.

Figure 4 illustrates the structure of the R-conv, which is a novel residual convolutional block that combines regular convolution with FConv in parallel to extract both global and

local features from gait sequences. The R-conv is designed to leverage the benefits of both global and local features by adding the output of FConv with the result of regular convolution. Assuming that the input is denoted as $X_{global} \in R^{c_1 \times t \times h \times w}$, where c_1 represents the number of channels, t means the length of feature maps and (h, w) indicate the dimensions of each image frame. Initially, the global feature map is horizontally partitioned into n parts to derive local feature maps, represented as $X_{local} = \left\{ X_{local}^{i} | i = 1, \cdots, n \right\}$, where n denotes the total number of partitions and $X_{local}^{i} \in R^{c_1 \times t \times \frac{h}{n} \times w}$ represents the i-th local gait segment. Subsequently, 2D CNN is used to extract global and local gait information independently. Finally, the combined features F_{R-conv}, which includes global and local information, are fused through element-wise summation, formulated as:

$$F_{R-conv} = F_{global} + F_{local} \in R^{c_2 \times t \times h \times w} \tag{2}$$

F_{global} and F_{local} can be expressed as:

$$F_{global} = F_{3\times3}\left(X_{global} \right) \in R^{c_2 \times t \times h \times w} \tag{3}$$

$$F_{local} = cat \begin{Bmatrix} F_{3\times3}\left(X_{local}^{1} \right) \\ F_{3\times3}\left(X_{local}^{2} \right) \\ \vdots \\ F_{3\times3}\left(X_{local}^{n} \right) \end{Bmatrix} \in R^{c_2 \times t \times h \times w} \tag{4}$$

where $F_{3\times3}$ represents the 2D convolution operation with the kernel size of (3,3), and cat denotes the concatenation operation on the horizontal dimension.

In FLSFE, except for three consecutive R-convs, SP is used to aggregate high-level feature maps from different gait timepoints that are robust to the appearance and observation perspectives into set-level feature maps. The set-level spatial representation is generated as $F_{SP} \in R^{C \times h \times w}$ and the process can be formulated as Equation (5):

$$F_{SP} = SP(F_{R-conv}) \tag{5}$$

Then, set-level features are summed up point by point and HP is designed to obtain the column vector spatial features $F_{space} \in R^{C_1 \times K}$ by dividing the feature map into K parts according to the height. The HP process can be formulated as Equation (6):

$$F_{spae} = maxpool\left(F_{sp} \right) + avgpool\left(F_{sp} \right) \tag{6}$$

where $maxpool$ represents the horizontal maximum pooling and $avgpool$ represents the horizontal average pooling.

3.2.3. Multi-Scale Temporal Feature Extraction (MSTFE)

MSTFE is devised to integrate long-term and short-term features in the temporal dimension, facilitating effective information exchange between different scales. As depicted in Figure 5, firstly, the HP operation is implemented on F_{MSFE} to obtain fine-grained information and then, on the one hand, the long-term features are extracted to represent the motion characteristics of all frames, which reveals the global motion periodicity of different body parts. On the other hand, two sequential 1D convolutions with a kernel size of 3 are used to extract short-term temporal contextual features that are favorable for modeling micromotion patterns. Finally, an adaptive temporal feature fusion approach is employed to acquire the temporal importance weights for each temporal scale at both long-term and short-term scales to adaptively highlight or suppress features to obtain the most discriminative motion characteristics.

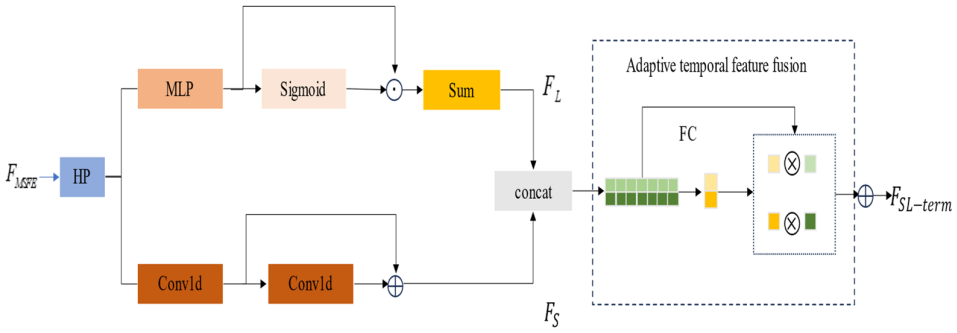

Figure 5. The detailed structure of multi-scale temporal feature extraction (MSTFE).

The input of MSTFE is denoted as F_{MSFE}, and the HP operation is used to partition the human body into K segments, resulting in the temporal feature representation $F_{Temp} \in R^{C\times T\times K}$, F^i_{Temp} denotes the i-th sample of each frame. Initially, a multilayer perceptron (MLP) followed by a Sigmoid function is applied to each frame to obtain the long-term temporal features. The MLP could be considered as a temporal attention mechanism, learning to assess the importance of different frames based on their relevance to the overall motion characteristics. Subsequently, the long-term temporal feature F_L is obtained by performing a weighted summation of all frames. The process can be expressed below:

$$F_L = \frac{\sum_{i=1}^{I} Sigmoid\left(MLP\left(F^i_{Temp}\right) \odot F^i_{Temp}\right)}{\sum_{i=1}^{I} Sigmoid\left(MLP\left(F^i_{Temp}\right)\right)}. \tag{7}$$

where $\odot$ denotes dot product and F_L describes the global motion cues.

To obtain the short-term temporal features, two sequential 1D convolutions with a kernel size of 3 are applied to the temporal feature representation F_{Temp} obtained from the HP operation. After each 1D convolution, the resulting features are summed to capture the short-term temporal feature F_S, which is formulated as:

$$F_s = Conv1d\left(F_{Temp}\right) \oplus Conv1d\left(Conv1d\left(F_{Temp}\right)\right) \tag{8}$$

To capture the varying motion patterns of different temporal scales and their varying importance in gait recognition, an adaptive temporal feature fusion approach is introduced. This approach aims to determine temporal importance weights that adaptively highlight or suppress features to obtain the most discriminative motion characteristics, which is realized through two fully connected layers followed by a Sigmoid function [35,36]. These layers are specifically designed to learn the temporal importance weights for both long-term and short-term temporal scales. The process can be expressed as follows:

$$W_T = Sigmoid(FC(FC(cat(F_s, F_L)))) \tag{9}$$

The fused temporal features can be realized by:

$$F^n_A = F^n_L \times W^n_{T,1} + F^n_S \times W^n_{T,2} \tag{10}$$

where W^n_T denotes the importance weight of the n-th frame of a sample, and the weights of the long-term and short-term temporal feature are denoted as $W^n_{T,1}$, $W^n_{T,2}$.

Based on the adaptive temporal feature fusion, we obtain sequence-level representations in a weighted summation way as follows:

$$F_{SL-term} = \frac{\sum_{n=1}^{N} F^n_A}{\sum_{n=1}^{N}\sum_{i-1}^{2} W^n_{T,i}} \tag{11}$$

where $F_{SL-term} \in R^{C_2 \times K}$, and C_2 means the number of channels.

3.3. View Embedding Module

Since view information is an imperative condition in gait recognition, the view embedding module is introduced to build view estimation and view angle is embedded into the previous two-path spatial-temporal feature fusion module, which can not only greatly minimize the intra-class caused by the view differences, but also improve the recognition ability of gait recognition.

3.3.1. View Prediction

As shown in Figure 6, F_{space} and $F_{SL-term}$ in the two-path parallel structure are concatenated along the channel dimension as the multi-scale spatial-temporal features F_{ST}, which can be formulated as:

$$F_{ST} = F_{spce} \Theta F_{SL-term} \tag{12}$$

where Θ denotes a merge operation on the channel dimension.

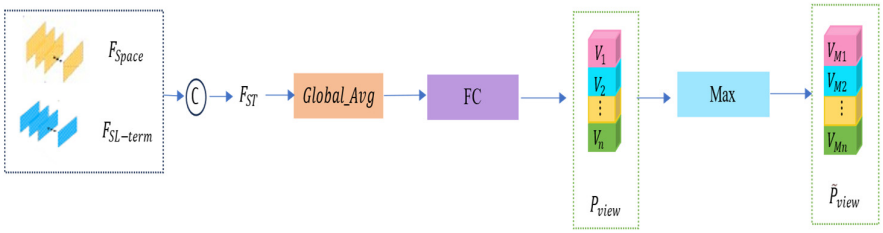

Figure 6. The detailed structure of view prediction.

The view classification feature can be realized by:

$$\tilde{F}_{view} = FC\left(P_{Global\_Avg}(F_{ST})\right) \tag{13}$$

where FC means a fully connect layer and P_{Global_Avg} indicates the global average pooling.

Subsequently, the view probability of the input silhouette, denoted as P_{view} and the maximum probability of the predicted view, represented as $\tilde{P}_{view}$ are determined by calculating the following formulas, respectively, as:

$$P_{view} = W_{view} \tilde{F}_{view} + B_{view} \tag{14}$$

$$\tilde{P}_{view} = Argmax(P_{view}) \tag{15}$$

where $P_{view} \in R^V$, V means the number of discrete views, W_{view} are the view weight matrix of FC layer, B_{view} denotes the bias terms of FC layer and $\tilde{P}_{view} \in \{0, 1, 2 \cdots V\}$.

Regarding the predicted discrete view $\tilde{P}_{view}$, a group of view projection matrix will be trained as $Z_{\tilde{P}_{view}} | = \{W_i | i = 1, 2, \cdots, n\}$, where $W_i \in R^{D \times D}$ is the projection matrix, and it will be used in Section 3.3.2.

3.3.2. HPP with View Embedding

In this section, horizontal pyramid pooling (HPP) [4] is utilized on the multi-scale spatial-temporal feature maps and then the weights under the best view are connected with these multi-scale spatial-temporal features by matrix multiplication to acquire the final feature.

The multi-scale spatial-temporal feature maps after HPP can be denoted as $F_{HPP,i}$, $i = 1, 2, 3, \cdots n$, where n means the number of strips into split in HPP, $F_{HPP,i} \in R^D$. Supposing the predicted view of F_{ST} is θ, the embed procedure can be realized by:

$$F_{final,i} = W_i F_{HPP,i} \tag{16}$$

$$F_{final} = \left[F_{final,1}, F_{final,2}, \cdots, F_{final,n}\right] \tag{17}$$

where $i = 1, 2, 3, \cdots n$, $W_i \in Z_\theta$, $F_{final,i} \in R^D$, $F_{final} \in R^{n \times D}$.

3.4. Joint Losses

We employ two joint loss functions, namely the cross entropy (CE) loss [37] and the triplet loss [38] for training our proposed framework. The CE loss is applied for the view prediction task and is calculated as follows:

$$L_{ce} = -\sum_{n=1}^{N}\sum_{V=1}^{V} y_n log(P_{n,v}) w.r.t. P_{n,v} = \frac{e^{\widetilde{P}_{viewn,v}}}{\sum_{V=1}^{V} e^{\widetilde{P}_{viewn,v}}} \tag{18}$$

where N represents the total number of silhouette sequences and y_n denotes the discrete ground value of the n-th sequence.

Next, the triplet loss is employed to augment the discriminative capability of the features. For each triplet of gait silhouette sequences (Q, P, N), where Q and P belong to the same subject, and Q and N are from the different subjects, the triplet loss can be calculated as follows:

$$L_{trip} = \frac{1}{K}\sum_{i=1}^{K}\sum_{l=1}^{n} max\left(m - d_{ij}^{-} + d_{ij}^{+},\ 0\right) d_{ij}^{-}. \tag{19}$$

where K represents the number of triplets, $d_{ij}^{-} = \left\| f_{final,j}^{Q_i} - f_{final,j}^{N_i} \right\|_2^2$ and $d_{ij}^{+} = \left\| f_{final,j}^{Q_i} - f_{final,j}^{P_i} \right\|_2^2$.

Finally, to obtain the overall loss function, we joint the CE loss and the triplet loss as follows:

$$L = L_{trip} + \lambda_{ce} L_{ce} \tag{20}$$

where λ_{ce} is a hyper-parameter.

4. Experiments

4.1. Datasets

CASIA-B: CASIA-B (http://www.cbsr.ia.ac.cn/, accessed on 28 August 2006) [39] is a well-known gait dataset with a total of 124 subjects. Each subject contains 11 views (from 0° to 180°). Each view includes 10 gait sequences captured under three various walking conditions: normal walking (NM) with 6 sequences, carrying bags (BG) with two sequences and wearing coats or jackets (CL) with two sequences. For our experiments, we follow the protocol from a previous work [20] and the dataset is split into a training set with samples from the initial 74 subjects, and a testing set with samples from the remaining subjects. During test, the first 4 sequences of the NM condition (NM #1–4) are kept as the gallery set, and the remaining 6 sequences (NM#5–6, BG#1–2, CL#1–2) are defined as the probe set, which aims to ensure consistency with the division of CASIA-B datasets in the state-of-the-art methods [4,7,12,21,29].

OU-MVLP: OU-MVLP (http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html, accessed on 4 October 2018) [40] is an extensive public gait dataset including 10,307 subjects in total. Among these subjects, 5153 are designated as training samples, while the remaining 5154 subjects are served as testing samples. Each subject has 14 distinct views (from 0° to 90° and 180° to 270°) and each view contains two sequences (#00–01).

4.2. Implementation Details

The batchsize is configured to (8, 8) and (32, 8) for CASIA-B dataset and OUMVLP dataset separately. The input silhouettes are resized to the size of 64 × 44 and aligned according to the method in [40]. Adam optimizer is applied for training with an initial learning rate of 0.0001 and the momentum 0.9 [41]. In the triplet loss L_{trip}, the margin is set to 0.2, and the λ_{ce} in Equation (20) is set to 0.05 and 0.3 on CASIA-B dataset and OUMVLP dataset separately. In CASIA-B dataset, three-layer CNNs are used in MSFE and three consecutive R-convs are implemented in FLSFE to extract features, while in OU-MVLP dataset, five-layer CNNs are applied in MSFE and an additional R-conv is stacked in FLSFE, and the value of n in each R-conv is set to 1, 1, 3, 3. Furthermore, the CASIA-B dataset is trained 70 K iteration, while the OUMVLP dataset is trained 250 K iteration. The learning rate is adjusted to 1×10^{-5} after 160 K iterations to ensure the stable convergence. All the experiments are carried out by using the Pytorch framework [42] on NVIDIA GeForce RTX 3090 GPUs [39].

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

CASIA-B: Table 1 presents the comprehensive comparison results between the proposed method and other state-of-the-art methods on CASIA-B dataset. From Table 1, it can be clearly seen that our method attains outstanding performance across nearly all viewpoints compared to other methods, such as GaitNet [12], GaitSet [4], GaitPart [7], MT3D [29], and RPNet [21]. In particular, the proposed method exhibits significantly higher average accuracies than other gait recognition methods, especially under BG and CL conditions. Our proposed method achieves average accuracies of 97.7%, 93.7%, and 83.8% under these conditions, outperforming GaitPart [7] by +1.5%, +2.2%, and +5.1%, respectively. Furthermore, our proposed method demonstrates superior performance in some specific view angles. For example, under the view angles of 90° and 180°, the NM accuracy of our method reaches 95.9% and 94.2%, surpassing GaitPart [7] by +3.4% and +3.8% separately. The experimental results illustrate that our framework exhibits substantial robustness and advantages under unfavorable conditions, which can be attributed to the fact that multi-scale convolution can simultaneously observe more detailed gait features in spatial and temporal domain. Furthermore, the effective combination of multi-scale spatial-temporal features can obtain discriminative features even in the presence of occlusion, thus enhancing the recognition ability of the proposed method.

OUMVLP: Similar experiments are executed on OU-MLVP dataset to prove the generalization efficacy of our approach. As demonstrated in Table 2, our method consistently attains higher accuracies across most camera views than other methods such as GaitSet [4], GaitPart [7] and RPNet [21], which displays a more outstanding recognition ability in the large-scale dataset. Specifically, the recognition ability of our method is superior to the state-of-the-art RPNet [18] by a margin of +4.4% (85.0% vs. 89.4%). Due to the addition of view embedding, the accuracies of our method under the view angles of 0° and 180° have been greatly improved. Meanwhile, it can be also observed that our results may not surpass GaitPart [7] under certain view angles, such as 45°, 75° and 90°, this is mainly due to the fact that our training dataset does not include as many samples as other works contained under these view angle. Although our results may not surpass GaitPart [7] under some specific view angles, our method has still achieved substantial overall improvement. In conclusion, our method can perform better recognition ability and effectively improve the recognition accuracy on the large and worldwide dataset for cross-view gait recognition.

Table 1. Averaged rank-1 recognition accuracies on CASIA-B dataset, excluding identical-view cases.

Gallery NM#1–4		0–180°											Mean
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5–6	GaitNet [12]	93.1	92.6	90.8	92.4	87.6	95.1	94.2	95.8	92.6	90.4	90.2	92.3
	GaitSet [4]	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 [7]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	MT3D [29]	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	RPNet [21]	95.1	99.0	99.1	98.3	95.7	93.6	95.9	98.3	98.6	97.7	90.8	96.6
	ours	**96.1**	**99.6**	**99.8**	**98.7**	**96.5**	**95.9**	**96.7**	**99.3**	**99.4**	**98.8**	**94.2**	**97.7**
BG#1–2	GaitNet [12]	88.8	88.7	88.7	94.3	85.4	92.7	91.1	92.6	84.9	84.4	86.7	88.9
	GaitSet [4]	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 [7]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	MT3D [29]	91.0	95.4	**97.5**	94.2	**92.3**	86.9	91.2	95.6	97.3	**96.4**	86.6	93.0
	RPNet [21]	92.6	92.3	96.6	94.5	91.9	87.6	90.7	94.7	96.0	93.9	86.1	92.8
	ours	**93.1**	**96.1**	97.2	**95.1**	91.6	**87.8**	**91.1**	**96.1**	**97.5**	95.8	**89.5**	**93.7**
CL#1–2	GaitNet [12]	50.1	60.7	72.4	72.7	74.6	78.4	70.3	68.2	53.5	44.1	40.8	62.3
	GaitSet [4]	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 [7]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	MT3D [29]	76.0	87.6	89.8	85.0	81.2	75.7	**81.0**	84.5	85.4	82.2	68.1	81.5
	RPNet [21]	75.6	87.1	88.3	87.1	83.1	78.0	79.9	82.7	83.9	78.9	66.6	80.3
	ours	**77.4**	**89.0**	**90.0**	**87.3**	**83.8**	**79.1**	80.7	**87.2**	**87.5**	**85.0**	**75.2**	**83.8**

Table 2. Averaged rank-1 recognition accuracies on OUMVLP dataset, excluding identical-view cases.

Probe	RPNet [21]	GaitSet [4]	GaitPart [7]	Ours
0°	73.5	79.5	82.6	**84.2**
15°	84.4	87.9	88.9	**89.9**
30°	89.6	89.9	90.8	**91.3**
45°	89.8	90.2	91.0	90.8
60°	86.3	88.1	89.7	**90.2**
75°	87.4	88.7	89.9	89.6
90°	86.0	87.8	89.5	88.9
180°	76.3	81.7	85.2	**86.7**
195°	83.2	86.7	88.1	**89.7**
210°	88.6	89.0	90.0	**90.5**
225°	88.9	89.3	90.1	**90.2**
240°	85.7	87.2	89.0	**89.7**
255°	86.4	87.8	89.1	**89.5**
270°	84.4	86.2	88.2	**89.8**
Mean	85.0	87.1	88.7	**89.4**

4.4. Ablation Study

In this paper, the three components (MSFE, FLSFE and MSTFE) of the two-path spatial-temporal feature fusion module and the view embedding module are included in the proposed framework. Hereby, exhaustive ablation experiments of these components are carried out on CASIA-B dataset to assess the effectiveness of each individual component. The experimental results and analysis are presented below.

Impact of Each component: Table 3 illustrates the effect of each component in our method on CASIA-B dataset. The baseline refers to the remaining structure after removing MSFE, FLSFE, MSTFE and view embedding module from the proposed methods, which

consists of three-layer CNNs and aggregates temporal information through the max operation. It can be observed from Table 3 that when MSFE is added to baseline alone, the average accuracy under three walking conditions achieves +1.5% improvement than the baseline, which implies that MSFE can perform well on the baseline by using different kernel sizes to handle the input. In addition, with the utilization of FLSFE and MSTFE, the average accuracy reaches 90.5%, which is +1.7% higher than GaitPart [7] and the accuracies under each condition are improved significantly, especially under BG and CL, proving the superiority of the two-path spatial-temporal feature fusion module under occlusion. Particularly, when FLSFE is added to the baseline, the average accuracy increases substantially by +7.9% compared to Baseline + MSFE, which verifies the fact that FLSFE plays a vitally important role in this two-path parallel structure. Finally, we incorporate view embedding into the baseline to increase the recognition accuracy of our method, resulting in a remarkable 91.7% performance, which is +1.2% higher than the previous version. The comparison results confirm that view prediction offers valuable perspective information for cross-view gait recognition and the view embedding module proves to be instrumental in improving the recognition capability of multi-view gait tasks.

Table 3. Study of the effectiveness of each component on CASIA-B dataset.

Model	NM	BG	CL	Mean
Baseline	89.8	82.0	59.5	77.1
Baseline + MSFE	91.6	83.5	60.7	78.6
Baseline + MSFE + FLSFE	95.6	88.6	75.2	86.5
Baseline + MSFE + FLSFE + MSTFE	97.0	92.9	81.6	90.5
Baseline + MSFE + FLSFE + MSTFE + view embedding	**97.7**	**93.7**	**83.8**	**91.7**

Impact of different kernel sizes in MSFE: To study the impact of various kernel sizes in MSFE, three various kinds of kernel sizes are designed and the ablation studies are conducted by arranging them. The results of these experiments are presented in Table 4. Evidently, increasing the number of multi-scale convolution kernels can attain higher recognition accuracies. Therefore, we set the kernel sizes of the multi-scale convolution to 1, 3, and 5, as it yields the most favorable recognition performance.

Table 4. Study of the effectiveness of different kernel sizes in MSFE on CASIA-B dataset.

Kernel Size	NM	BG	CL
1	89.8	82	59.5
1, 3	91.3	82.8	60.2
1, 5	90.9	82.6	59.9
1, 3, 5	**91.6**	**83.5**	**60.7**

Impact of the value of *n* in R-conv: Following the approach of setting the parameter n in R-conv as described in Section 3.2.2, five controlled experiments are performed and the results are presented in Table 5. Notably, when the n value of all the R-conv is set to 1, the R-conv becomes entirely composed of regular layers. From Table 5. it becomes evident that when the n value of three consecutive R-conv is set to 2, 8, 8, the recognition accuracy achieves the most excellent performance compared to other four experiments. There is another thing worth noting that it is not the case that the larger n is, the higher recognition accuracy will be. For example, when the n value of three consecutive R-conv is set to 4, 8, 8, the accuracy under NM increases slightly, while the accuracies under BG and CL decrease dramatically.

Table 5. Study of the effectiveness of the value of n in each R-conv on CASIA-B dataset.

R-conv1	R-conv2	R-conv3	NM	BG	CL
1	1	1	95.3	86.1	71.9
2	2	2	95.3	86.9	72.8
2	4	4	95.4	87.8	73.9
2	8	8	95.6	**88.6**	**75.2**
4	8	8	**95.9**	87.9	74.7

Impact of multi-scale temporal features: We conduct an investigation into the effects of the multi-scale temporal features in MSTFE and the results are presented in Table 6 Obviously, both the long-term and short-term temporal features produce positive effects on recognition performance, and joining these two muti-scale temporal features achieves the best performance, since the long-term and short-term features can interact with each other, increasing the diversity of temporal representations and further improving the overall recognition accuracy.

Table 6. Study of the effectiveness of multi-scale temporal features on CASIA-B dataset.

Short-Term	Long-Term	NM	BG	CL
✓		96.8	91.9	80.6
	✓	95.5	90.3	74.3
✓	✓	**97.0**	**92.9**	**81.6**

5. Discussion

Based on extensive comparative experimental results, our proposed model exhibits significant improvements in two key aspects: (1) Enhanced accuracy under BG and CL conditions on CASIA-B dataset. This improvement is attributed to the utilization of MSFE, FLSFE and MSTFE. MSFE can expand different perceptual fields and observe more detailed gait features in spatial and temporal domain. Besides, both FLSFE and MSTFE in a two-path parallel structure can extract muti-scale spatial and temporal discriminative features, which can enhance the robustness of the proposed method, particularly under unfavorable conditions. (2) Enhanced accuracy on the OUMVLP dataset. The proposed method demonstrates outstanding performance in the large-scale dataset, this is due to the fact that the view embedding module can predict the best view and embed view angle into the multi-scale spatial-temporal features, which can help mitigate the intra-class variations resulting from view differences and enhance the recognition ability of gait recognition.

In the future, we will further improve the performance of our proposed method in more complex test scenarios and gait in-the-wild datasets, such as the GREW [43] dataset and the Gait3D [44] dataset. Simultaneously, whether our method exists overfitting over CASIA-B and OU-MVLP datasets also needs to be verified in the wild datasets and real-word scenarios. Additionally, as silhouettes are easily disturbed by pedestrians' clothes and objects, it is essential to explore multi-modal gait recognition approaches that combines silhouettes, skeletons [45], and pose heatmaps [46].

6. Conclusions

In this paper, we propose a novel gait recognition framework that combines two-path spatial-temporal feature fusion with view embedding. In the two-path spatial-temporal feature fusion module, MSFE is firstly utilized to extract feature representations with different granularities from the input frames. Then, the two-path parallel structure is designed to obtain the muti-scale spatial-temporal features, where FLSFE extracts both global and local features by R-convs and MSTFE interacts temporal features with multiple scales for achieving the strong ability of muti-scale spatial-temporal modeling. Additionally, a view embedding module is put forward to make the muti-scale spatial-temporal features

under the best viewpoint. Extensive experiments conducted on various public datasets validate superior performance of our approach compared to recent counterparts.

Author Contributions: Conceptualization, C.H.; Data curation, D.G. and X.Z.; Formal analysis, C.H. and D.G.; Funding acquisition, C.H.; Investigation, D.G.; Methodology, C.H. and D.G.; Project administration, C.H.; Resources, C.H.; Software, D.G.; Supervision, C.H.; Validation, C.H. and D.G.; Visualization, D.G. and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the Research Foundation of Education Bureau of Liaoning Province (Grant No. LZD202001) and the Science and Technology Project of Department of Science & Technology of Liaoning Province (Grant No. 2021JH1/10400029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request. The data are not publicly available due to personal information and privacy about the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, J.; Zheng, N. Gait History Image: A Novel Temporal Template for Gait Recognition. In Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, Beijing, China, 2–5 July 2007.
2. Singh, S.; Biswas, K.K. Biometric Gait Recognition with Carrying and Clothing Variants. Pattern Recognition and Machine Intelligence. In Proceedings of the Third International Conference, New Delhi, India, 16–20 December 2009.
3. Huang, S.; Elgammal, A.; Lu, J.; Yang, D. Cross-speed Gait Recognition Using Speed-Invariant Gait Templates and Globality–Locality Preserving Projections. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 2071–2083. [CrossRef]
4. Chao, H.; He, Y.; Zhang, J.; Feng, J. Gaitset: Regarding Gait as A Set for Cross-View Gait Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
5. Hou, S.; Cao, C.; Liu, X.; Huang, Y. Gait Lateral Network: Learning Discriminative and Compact Representations for Gait Recognition. In Proceedings of the European Conference on Computer Vision, Edinburgh, UK, 23–28 August 2020.
6. Chen, Y.; Zhao, Y.; Li, X. Spatio-Temporal Gait Feature with Adaptive Distance Alignment. *arXiv* **2022**, arXiv:2203.03376v3.
7. Fan, C.; Peng, Y.; Cao, C.; Liu, X. Gaitpart: Temporal Part-Based Model for Gait Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–20 June 2020.
8. Huang, Z.; Xue, D.; Shen, X.; Tian, X. 3D Local Convolutional Neural Networks for Gait Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
9. Liao, R.; Cao, C.; Garcia, E.B.; Yu, S. Pose-based Temporal-Spatial Network (PTSN) for Gait Recognition with Carrying and Clothing Variations. In Proceedings of the Chinese Conference on Biometric Recognition, Shenzhen, China, 28–29 October 2017.
10. Liao, R.; Yu, S.; An, W.; Huang, Y. A Model-based Gait Recognition Method with Body Pose and Human Prior Knowledge. *Pattern Recognit.* **2020**, *98*, 107069. [CrossRef]
11. Shiraga, K.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Geinet: View-Invariant Gait Recognition Using a Convolutional Neural Network. In Proceedings of the 2016 International Conference on Biometrics, Halmstad, Sweden, 13–16 June 2016.
12. Memory, L.S.T. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780.
13. Wang, Y.; Sun, J.; Li, J.; Zhao, D. Gait Recognition Based on 3D Skeleton Joints Captured by Kinect. In Proceedings of the 2016 IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016.
14. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
15. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
16. Guler, R.A.; Neverova, N.; DensePose, I.K. Densepose: Dense Human Pose Estimation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
17. Ben, X.; Zhang, P.; Lai, Z.; Yan, R.; Zhai, X. A General Tensor Representation Framework for Cross-View Gait Recognition. *Pattern Recognit.* **2019**, *90*, 87–98. [CrossRef]
18. Fan, C.; Liang, J.; Shen, C.; Hou, S.; Huang, Y. OpenGait: Revisiting Gait Recognition Towards Better Practicality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
19. Zhang, C.; Liu, W.; Ma, H.; Fu, H. Siamese Neural Network Based Gait Recognition for Human Identification. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, China, 20–25 March 2016.
20. Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T. A Comprehensive Study on Cross-view Gait Based Human Identification with Deep CNNs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 209–226. [CrossRef] [PubMed]

21. Zhang, Y.; Huang, Y.; Yu, S.; Wang, L. Cross-view Gait Recognition by Discriminative Feature Learning. *IEEE Trans. Image Process.* **2019**, *29*, 1001–1015. [CrossRef] [PubMed]
22. Qin, H.; Chen, Z.; Guo, Q.; Wu, Q.J.; Lu, M. RPNet: Gait Recognition with Relationships Between Each Body-Parts. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2990–3000. [CrossRef]
23. Lin, B.; Zhang, S.; Yu, X. Gait Recognition via Effective Global-Local Feature Representation and Local Temporal Aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Beijing, China, 20–25 June 2021.
24. Wang, M.; Lin, B.; Guo, X.; Li, L.; Zhu, Z. GaitStrip: Gait Recognition via Effective Strip-Based Feature Representations and Multi-Level Framework. In Proceedings of the Asian Conference on Computer Vision, Macau, China, 4–8 December 2022.
25. Huang, X.; Zhu, D.; Wang, H.; Wang, X.; Yang, B. Context-Sensitive Temporal Feature Learning for Gait Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
26. Wu, H.; Tian, J.; Fu, Y.; Li, B.; Li, X. Condition-Aware Comparison Scheme for Gait Recognition. *IEEE Trans. Image Process.* **2020**, *30*, 2734–2744. [CrossRef] [PubMed]
27. Zhang, Z.; Tran, L.; Yin, X.; Atoum, Y.; Liu, X.; Wan, J.; Wang, N. Gait Recognition via Disentangled Representation Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
28. Ariyanto, G.; Nixon, M.S. Marionette Mass-spring Model for 3D Gait Biometrics. In Proceedings of the 2012 5th IAPR International Conference on Biometrics, New Delhi, India, 29 March–1 April 2012.
29. Ariyanto, G.; Nixon, M.S. Model-Based 3D Gait Biometrics. In Proceedings of the 2011 International Joint Conference on Biometrics, Washington, DC, USA, 11–13 October 2011.
30. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
31. Wolf, T.; Babaee, M.; Rigoll, G. Multi-View Gait Recognition Using 3D Convolutional Neural Networks. In Proceedings of the 2016 IEEE International Conference on Image Processing, Phoenix, AZ, USA, 25–28 September 2016.
32. Lin, B.; Zhang, S.; Bao, F. Gait Recognition with Multiple-Temporal-Scale 3D Convolutional Neural Network. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020.
33. He, Y.; Zhang, J.; Shan, H.; Wang, L. Multi-Task GANs for View-Specific Feature Learning in Gait Recognition. *IEEE Trans. Inf. Forensics Secur.* **2018**, *14*, 102–113. [CrossRef]
34. Chai, T.; Mei, X.; Li, A.; Wang, Y. Silhouette-Based View-Embeddings for Gait Recognition under Multiple Views. In Proceedings of the IEEE International Conference on Image Processing, Anchorage, AK, USA, 19–22 September 2021.
35. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B. Recent Advances in Convolutional Neural Networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
36. Huo, Y.; Gang, S.; Guan, C. FCIHMRT: Feature Cross-Layer Interaction Hybrid Method Based on Res2Net and Transformer for Remote Sensing Scene Classification. *Electronics* **2023**, *12*, 4362. [CrossRef]
37. Zhang, Z.; Sabuncu, M. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3 December 2018.
38. Hermans, A.; Beyer, L.; Leibe, B. In Defense of The Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.
39. Yu, S.; Tan, D.; Tan, T. A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. In Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 20–24 August 2006.
40. Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Multi-View Large Population Gait Dataset and Its Performance Evaluation for Cross-View Gait Recognition. *IPSJ Trans. Comput. Vis. Appl.* **2018**, *10*, 4. [CrossRef]
41. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
42. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E. Automatic Differentiation in Pytorch. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
43. Zhu, Z.; Guo, X.; Yang, T.; Huang, J.; Deng, J. Gait Recognition in the Wild: A Benchmark. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
44. Zheng, J.; Liu, X.; Liu, W.; He, L.; Yan, C. 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, New Orleans, LA, USA, 18–24 June 2022.
45. Yao, L.; Kusakunniran, W.; Wu, Q.; Xu, J.; Zhang, J. Collaborative Feature Learning for Gait Recognition under Cloth Changes. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 3615–3629. [CrossRef]
46. Zhao, L.; Guo, L.; Zhang, R.; Xie, X.; Ye, X. MmGaitSet: Multimodal Based Gait Recognition for Countering Carrying and Clothing Changes. *Appl. Intell.* **2022**, *52*, 2023–2036. [CrossRef]