electronics

MDPI

Article

Metric-Based Key Frame Extraction for Gait Recognition

Tuanjie Wei [1], Rui Li [1,2,*], Huimin Zhao [1,*], Rongjun Chen [1], Jin Zhan [1], Huakang Li [3] and Jiwei Wan [1]

[1] School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou 510665, China
[2] School of Art Design, Guangzhou College of Commerce, Guangzhou 511363, China
[3] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
* Correspondence: lirui198512@126.com (R.L.); zhaohuimin@gpnu.edu.cn (H.Z.)

Abstract: Gait recognition is one of the most promising biometric technologies that can identify individuals at a long distance. From observation, we find that there are differences in the length of the gait cycle and the quality of each frame in the sequence. In this paper, we propose a novel gait recognition framework to analyze human gait. On the one hand, we designed the Multi-scale Temporal Aggregation (MTA) module that models temporal and aggregate contextual information with different scales, on the other hand, we introduce the Metric-based Frame Attention Mechanism (MFAM) to re-weight each frame by the importance score, which calculates using the distance between frame-level features and sequence-level features. We evaluate our model on two of the most popular public datasets, CASIA-B and OU-MVLP. For normal walking, the rank-1 accuracies on the two datasets are 97.6% and 90.1%, respectively. In complex scenarios, the proposed method achieves accuracies of 94.8% and 84.9% on CASIA-B under bag-carrying and coat-wearing walking conditions. The results show that our method achieves the top level among state-of-the-art methods.

Keywords: gait recognition; temporal modeling; key frame; frame attention mechanism

Citation: Wei, T.; Li, R.; Zhao, H.; Chen, R.; Zhan, J.; Li, H.; Wan, J. Metric-Based Key Frame Extraction for Gait Recognition. *Electronics* **2022**, *11*, 4177. https://doi.org/10.3390/electronics11244177

Academic Editor: Hyunjin Park

Received: 11 November 2022
Accepted: 10 December 2022
Published: 14 December 2022

1. Introduction

Nowadays, gait recognition plays a significant role in personal identification, and different from other biometrics such as the face, fingerprint, and iris, the human gait is the only one that can be captured in long-distance conditions and the recognition process does not need the subject's cooperation. Therefore, with the popularization of video surveillance equipment, gait recognition technology has broad applications in crime prevention, forensic identification, and social security. However, in real-world scenarios, the performance of gait recognition suffers from many conditions such as changing clothing, carrying conditions, and the camera's viewpoint.

Recently, lots of deep convolutional neural-network-based methods have been proposed to address these issues. Zhang et al. [1] proposed a new auto-encoder framework to explicitly separate posture and appearance features from RGB images and then used LSTMs to model the temporal changes of gait sequence. Chao et al. [2] hypothesized that the appearance of a silhouette contains position information and the sequence information of gait was unnecessary for recognition, so they proposed a novel network named GaitSet that regarded gait silhouettes as a set to extract temporal information. Fan et al. [3] employed partial features for a human body description and proposed a new model named GaitPart, which focuses on the short-range temporal features rather than the redundant long-range features for gait cycles. Li et al. [4] enhanced the fine-grained learning of human partial features by segmenting and associating adjacent body parts from top to bottom. Lin et al. [5] assumed that the representations based on global information often neglect the details of the gait frame, while local region-based descriptors cannot capture the relations among neighboring regions, and thus designed a new global and local feature extraction module to address this issue.

These previous methods [3,4] extract fine-grained features from human body parts and model short-term motion patterns, effectively improving the performance of gait recognition models. However, from observation, we find that the differences in the pedestrian walking speed and camera frame rate have resulted in inconsistent frame length of the gait cycle (as shown in Figure 1), and thus the single temporal modeling approach cannot adapt to the diversity of motion.

Figure 1. Two complete gait cycle sequences from subjects '01' and '10' on CASIA-B. The top periodic sequence contains 24 frames, while the bottom contains 30 frames; the quality of frames in the sequence is different, and the frames with clear outlines and rich information are regarded as key frames (such as the frames in the red box).

Furthermore, unlike face recognition [6,7], fingerprint recognition [8,9], etc., which extracts identity information from a single image, gait recognition technology is based on video sequences. In the early days, most methods [10–12] fused sequence features by generating template images. In recent years, video sequence-based methods [2–5] aggregate sequence-level features by simple temporal pooling of frame-level features; however, this adaptive fusion method ignores the differences in the quality of frames (as shown in Figure 1), which affects the performance of the gait recognition model in various scenarios.

To alleviate these issues, we propose a novel gait recognition framework, which consists of two well-designed novel components, namely Multi-scale Temporal Aggregation (MTA) and Metric-based Frame Attention Mechanism (MFAM). MTA aggregates multi-scale context information through gait temporal modeling. MFAM calculates the importance score of each frame according to the Euclidean distance between it and the aggregated sequence features.

In summary, the major works of this paper are as follows:

(1) We propose the Multi-scale Temporal Aggregation module, which models gait temporal information in multiple scales, to accommodate diverse representations of motion;
(2) We introduce the Metric-based Frame Attention Mechanism, which assigns weights to each frame with an importance score calculated by the distance between frame-level features and sequence-level features;
(3) The proposed method has been evaluated on the widely used CASIA-B [13] and OU-MVLP [14] gait benchmark datasets. The experimental results of our method achieve high recognition accuracies under cross-view and various walking conditions.

2. Related Work

2.1. Gait Recognition

Existing gait recognition methods can be divided into two main categories: model-based [15,16] and sequence-based [17–20]. Model-based methods use the structure or motion model of the human body, such as gait period, stride scale, joint angle trajectories, etc. Liao et al. [21]. used the pose estimation method to extract 2D pose key points and extracted spatiotemporal invariant features from the gait pose, which effectively improved the performance of the network. Later, Liao et al. [22] assumed that the 3D pose defined by 3D coordinates was constant, and combined them with human pose priors, such as the motion relationship of the upper and lower limbs, motion trajectories, etc., to extract gait features, which improved the accuracy of the algorithm in the scene of changing perspectives. Model-based methods are not sensitive to changes in covariates such as viewing angle and clothing; however, the model-based method's recognition accuracy depends on the performance of pose estimation algorithms [23,24].

Early sequence-based methods usually compress the frame sequence into a gait template image (such as Gait Energy Image (GEI) [10], gait entropy image (GEnI) [11], period energy image (PEI) [12], etc.), and then extract gait features from the template image. The similarity between features is measured by machine learning algorithms, and finally, a label is assigned to each template image by some classifiers. Recently, due to the good performance of deep learning in various image processing tasks, Thomas et al. [25] applied 3D CNN to capture robust spatial–temporal gait features in multiple views; however, traditional 3D CNNs require fixed-length gait sequences for classification and thus are not able to address different lengths of videos directly. Chao et al. [2] regarded the gait silhouette sequence as a set and propose a new network named GaitSet to learn identity information from the set. The GaitSet model improves the recognition rate in various scenarios and ensures flexibility and effectiveness. Hajra et al. [26] utilize gait dynamics for gait feature extraction and the spatiotemporal power spectral gait features are utilized for a quadratic support vector machine classifier for gait recognition.

2.2. Temporal Modeling

In the literature, GaitSet considered a gait sequence as an unordered set consisting of independent frames, which ensures the flexibility of the model but limits the application efficiency of temporal information. To extract temporal features of gait, 1D temporal convolution and LSTM are usually used for gait temporal modeling. Fan et al. [3] believed that each part of the human body has its own unique motion pattern, and uses one-dimensional convolution to extract local short-range temporal features. Zhang et al. [1] used an LSTM network to achieve long-short temporal modeling of gait. Lin et al. [5] assumed that the set pooling operation [2,3] will bring the loss of spatial information, and thus proposed a novel local temporal aggregation operation to aggregate local temporal information.

The LSTM-based method [1,19] preserves unnecessary temporal constraints and is computationally expensive. The short-term modeling method improves recognition accuracy to a certain extent. However, the single-time modeling method cannot adapt to the complexity of motion and the change of realistic factors. Therefore, we propose the Multi-scale Temporal Aggregation module to aggregate temporal features from multiple different scales, so that the model can adapt to the diversity of motion.

2.3. Key Frame in Sequence

Compared with images, videos are richer in spatiotemporal information. However, there is too much redundant information in sequence, so extracting the information of key frames is crucial for many tasks. In the person re-identification task, Song et al. [27] propose an RQEN model, which judges the quality of pictures and reduces the importance of poor-quality frames. Ding et al. [28] propose a new key frame extraction method, frame difference and cluster (FDC), that integrates the idea of K-means clustering. To obtain more discriminative gait features, Wang et al. [29] propose a feature extraction algorithm based on local gait energy image (LGEI) and calculate LGEIs for each key frame. Li et al. [4] propose a residual frame attention mechanism (RFAM) module to highlight the key frames of sequences based on the slice features to extract the key frames from each body part.

Current gait recognition methods usually aggregate frame-level features into sequence features with adaptive temporal pooling, which ignores the importance of differences between frames. Therefore, we propose the Metric-based Frame Attention Mechanism (MFAM) module, which scores the importance of each frame by measuring the distance between sequence-level features and frame-level features, and then generates an appropriate sequence-level representation in a weighted summation manner to highlight key frames within the sequence.

3. Proposed Method

In this section, we first overview the framework of the proposed method. Then introduce the Multi-scale Temporal Aggregation (MTA) and the Metric-based Frame At-

tention Mechanism (MFAM). Finally, we introduce the details of training and testing. The framework of the proposed algorithm is shown in Figure 2.

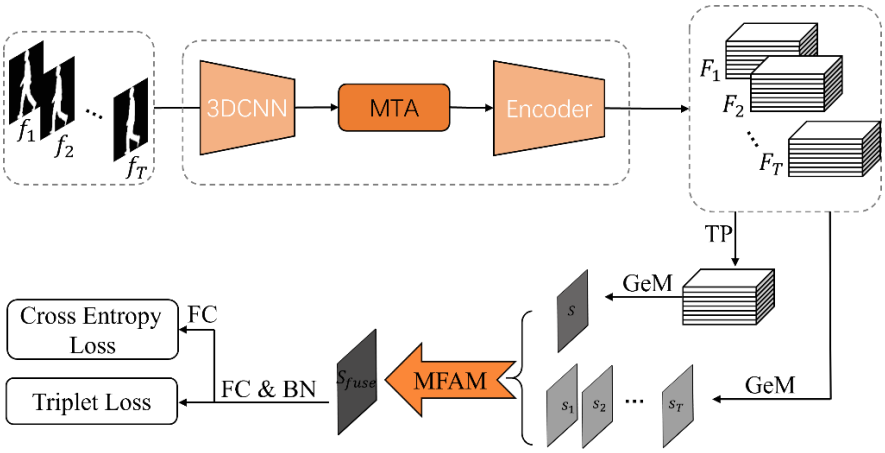

Figure 2. Overview of the gait recognition framework. The Multi-scale Temporal Aggregation (MTA) module extracts contextual information at different temporal scales. The Metric-based Frame Attention Mechanism (MFAM) is employed to highlight key frames in sequences. The TP and GeM represent the Temporal Pooling and Generalized-Mean pooling, respectively.

3.1. Overview

As shown in Figure 2, the input of the framework is a sequence of gait silhouettes, which has a dimension of $C * T * H * W$, where C represents the number of channels for the input frame, T represents the frame number of sequences, and H, W represent the height and width dimensions, respectively. In the framework, a 3D convolution is used to extract shallow features from the original input sequence, and the extracted shallow features dimensions are $C_1 * T * H * W$. Then, the Multi-scale Temporal Aggregation (MTA) module is designed with several parallel temporal convolutional layers, which aggregate temporal features from multiple different scales. The output dimensions of the MTA module are $C_1 * T_1 * H * W$. After that, the encoder network is used to extract frame-level features which are denoted as $F_t(t \in 1, 2, \ldots, T_1)$ with the dimension of $C_2 * H_1 * W_1$. In this paper, we adopt the GLFE proposed by Lin et al. [5] as the encoder. Specifically, the encoder consists of a global feature extraction branch and a local feature extraction branch, each containing 3 convolutional layers. There is a pooling layer after the first convolutional layer. The features of the two branches are fused using an addition or concatenation operation, and the concatenation operation is only performed after the last layer of convolution.

Then, we employ two parallel branches to process frame-level features separately. On the one hand, temporal pooling (TP) operation is adopted to aggregate frame-level features into a sequence-level feature which is denoted as F and with the dimension of $C_2 * H_1 * W_1$. Similar operations were commonly used in [2–4]. Then, in order to reduce the redundancy of data, the Generalized-Mean pooling (GeM) [5] operation is used to map the sequence-level feature (F) into 1D feature vector (denoted as S with the dimension of $C_2 * K$). On the other hand, the GeM operations are also used directly to reduce data redundancy for frame-level features.

Finally, the Metric-based Frame Attention Mechanism (MFAM) is designed to calculate the importance score for each frame and thus re-weight for all frames; the weighted features are encoded into high-dimensional vectors using several separate FC layers as gait representations.

3.2. Multi-Scale Temporal Aggregation

As discussed in Section 2.2, existing methods [2–4] either only focus on spatial modeling and thus ignore the inter-frame dependence, or focus on the short-term features of gait

cycles, which cannot adapt to the changes of complex motion and environmental factors. Therefore, we designed the Multi-scale Temporal Aggregation (MTA) module, which aims at aggregating contextual information at different scales.

As shown in Figure 3, the input of MTA module has a dimension of $C_1 * T * H * W$, which represents the number of channels, sequence length, and size of each frame, respectively. In order to capture the temporal features with different scales, which are denoted as T_s, T_m, and T_l, respectively, MTA employs three parallel convolutions, which adopt different sizes of the kernel. The specific parameter settings of the convolutional layers are shown in Table 1, especially, the three convolutions stride are all three.

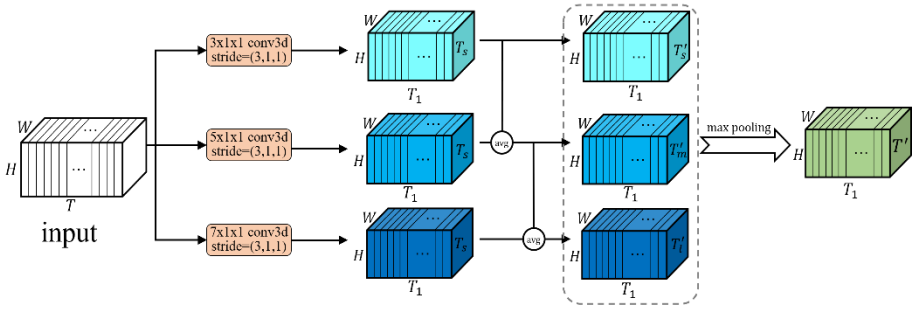

Figure 3. The structure of Multi-scale Temporal Aggregation (MTA). The MTA module employed several parallel convolutions and slides on sequence dimensions to obtain the features at different scales.

Table 1. The structure of MTA.

	kernel_1	**kernel_2**	**kernel_3**
kernel_size	(3,1,1)	(5,1,1)	(7,1,1)
padding	(0,0,0)	(1,0,0)	(2,0,0)

After that, MTA applies information flowing from small scale to large scale among temporal features by average pooling (*avg*), and the formula is as follows:

$$\begin{aligned} T'_s &= T_s \\ T'_m &= avg(T_s, T_m) \\ T'_l &= avg(T_s, T_m, T_l) \end{aligned} \tag{1}$$

Finally, MTA adopts max pooling (*max*) to aggregate context information at different scales.

$$T' = max\left(T'_s,\ T'_m, T'_l\right) \tag{2}$$

Based on the MTA module, the model aggregates context information of different time scales, provides multiple time receptive fields through information exchange and fusion between features, and can effectively adapt to the diverse expression of human motion.

3.3. Metric-Based Frame Attention Mechanism

The rich spatiotemporal features in video sequences provide more discriminative information for recognition tasks; however, redundant information in videos will also affect recognition accuracy. Therefore, researchers [4,27–29] focus on extracting key information from videos to improve the performance of the model. In this paper, we obtained the importance score of each frame by calculating the similarity between frame-level features and sequence-level features obtained by temporal pooling, and then weighted each frame and aggregated it into sequence features.

As shown in Figure 4, the input of MFAM module can divide into a frame-level feature vector and sequence-level feature vector, with dimensions of $C_2 * T_1 * K$ and $C_2 * K$,

respectively, where C_2 represents the number of channels, T_2 represents the sequence length, and K represents the number of preset body parts.

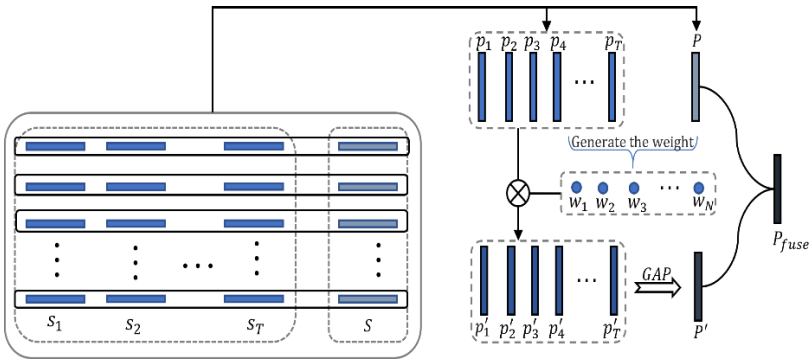

Figure 4. The structure of Metric-based Frame Attention Mechanism (MFAM). p_t and P represent a set of human body parts feature vectors with frame-level features (s_t) and sequence-level features (S), respectively.

In order to assess the importance of each frame, MFAM first calculates the Euclidean distance between frame-level feature vectors and sequence-level feature vectors and uses max-min normalized processing of the numeric distance. Then, we take the negative of these numeric distances and apply the sigmoid activation function to capture the score (denoted as w_t) of each frame. The score of each frame is calculated as follows:

$$\begin{aligned} D_t &= ||p_t - P||_2 \\ w_t &= \sigma\left(-\frac{D_t - \min(D_t)}{\max(D_t) - \min(D_t)}\right) \end{aligned} \tag{3}$$

where σ represents the sigmoid function and min and max represent the minimum and maximum scores, respectively.

After that, all frames are weighted by the importance score, and the re-weight frame-level features (denoted as p'_t) are aggregated into re-weight sequence-level features (denoted as P') by temporal pooling (TP). In this paper, temporal pooling module employs global max pooling (GAP) operation to aggregate features. The calculation processes are as follows:

$$\begin{aligned} p'_t &= w_t * p_t \\ P' &= GAP(p'_t) \end{aligned} \tag{4}$$

Finally, MFAM fuses the initial sequence-level features with weighted sequence-level features as the final sequence-level features (denoted as P_{fuse}).

$$P_{fuse} = P + P' \tag{5}$$

3.4. Training and Testing

During the training stage, we input a gait sequence into the network and obtained the gait feature descriptors; the batch size of the input training data is $p * k$, where p represents the number of persons and k represents the number of training samples of each person in the batch. Then, the Batch All (BA+) triplet loss [30] and cross-entropy loss are employed to optimize the model.

During the Testing Stage, the test dataset is divided into gallery set and probe set, and we input the whole gait sequences into the network to generate gait feature descriptors. To calculate rank-1 accuracy, the gallery set is regarded as the standard view to be retrieved, and the descriptors of the probe are used to match the descriptors from the gallery view based on the average Euclidean distance.

4. Experimental Results

4.1. Dataset

We use two open databases, CASIA-B [13] and OU-MVLP [14], to evaluate the performance of the proposed method.

CASIA-B. Contains the gait sequences of 124 subjects, CASIA-B is a widely applied gait dataset, and each subject contains 3 walking conditions and 11 views (0°~180°, with an interval of 18°). The walking condition contains normal (NM) (six sequences per subject), walking with a bag (BG) (two sequences per subject), and wearing a coat or jacket (CL) (two sequences per subject). In other words, each subject contains $11 \times (6 + 2 + 2) = 110$ sequences. As there is no official partition of training and test sets of this dataset, we conduct large-sample training (LT), medium-sample training (MT), and small-sample training (ST) according to Chao et al. [2] During the testing stage, the first four sequences of the NM condition (NM#1–4) are stored in the gallery set and the remaining six sequences (NM#5–6, BG#1–2 and Cl#1–2) are stored in the probe set.

OU-MVLP. Contains the gait sequences of 10307 subjects, OU-MVLP is so far the world's largest public gait dataset, and each subject contains two sequences (#00 and #01), with fourteen views (0°, 15°, ... , 90°, 180°, 195°, ... , 270°) for each sequence. The sequences are divided into training and test sets by subjects (5153 subjects for training and 5154 subjects for testing). During the testing stage, sequences with index #01 are kept in a gallery and those with index #00 are used as probes.

4.2. Training and Testing Details

In all the experiments, the silhouettes are directly provided by the datasets and are aligned by the method proposed by Takemura et al. [14] and resized to the size of 64×44. In the training stage, we choose Adam [31] as an optimizer and set the margin in Batch All (BA+) triplet loss to 0.2. For CASIA-B, the batch size parameters P and K are set to 8 and 10. In the setting of ST, MT, and LT, the epoch number is set to 60 K, 80 K, and 80 K, respectively, and the learning rate is set to 1×10^{-4}. For OUMVLP, the batch size parameters P and K are set to 32 and 10, respectively, and the epoch number is set to 250 K. The learning rate is first set to 1×10^{-4}, and reset to 1×10^{-5} and 5×10^{-6} after 180 K and 230 K iterations, respectively.

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

4.3.1. Experimental Results on CASIA-B Dataset

In order to verify the efficacy and superiority of our method, we compare the performance of our model with other state-of-the-art gait recognition models, including CNN-LB [32], MGAN [33], GaitSet [2], GaitSlice [4], and GaitGL [5] on the CASIA-B gait dataset based on the rank-1 accuracy. The experimental results are shown in Tables 2–4, and Figure 5. Except for ours, other results are directly taken from their original papers. All the results are averaged on the 11 gallery views and the identical views are excluded.

Table 2. Averaged rank-1 accuracies on CASIA-B under LT setting, 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	CNN-LB	82.6	90.3	96.1	94.3	90.1	87.4	89.9	94.0	94.7	91.3	78.5	89.9
	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitSlice	95.5	99.2	99.6	99.0	94.4	92.5	95.0	98.1	99.7	98.3	92.9	96.7
	GaitGL	96.0	98.3	**99.0**	97.9	**96.9**	**95.4**	97.0	**98.9**	**99.3**	**98.8**	94.0	97.4
	Ours	**96.7**	**98.6**	98.9	**98.0**	96.7	**95.4**	**97.2**	98.7	99.2	98.7	**95.2**	**97.6**
BG#1-2	CNN-LB	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitSlice	90.2	96.4	96.1	94.9	89.3	85.0	90.9	94.5	96.3	95.0	88.1	92.4
	GaitGL	92.6	**96.6**	**96.8**	95.5	**93.5**	89.3	92.2	96.5	**98.2**	**96.9**	91.5	94.5
	Ours	**93.9**	96.5	**96.8**	**95.9**	**93.5**	**89.6**	**92.5**	**97.0**	98.0	96.8	**91.9**	**94.8**
CL#1-2	CNN-LB	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitSlice	75.6	87.0	88.9	86.5	80.5	77.5	79.1	84.0	84.8	83.6	70.1	81.6
	GaitGL	76.6	**90.0**	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	Ours	**78.0**	**90.0**	**91.6**	**88.5**	**84.6**	**79.8**	**84.7**	**88.2**	**88.2**	**86.4**	**73.4**	**84.9**

Table 3. Averaged rank-1 accuracies on CASIA-B under MT setting, 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	MGAN	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	76.2	68.6	53.8	68.1
	GaitSet	86.8	95.2	98.0	94.5	91.5	89.1	91.1	95.0	97.4	93.7	80.2	92.0
	GaitSlice	92.2	97.3	98.9	98.4	94.2	90.3	94.2	97.5	99.2	96.6	89.4	95.3
	GaitGL	93.9	97.6	98.8	**97.3**	95.2	**92.7**	95.6	98.1	98.5	96.5	91.2	95.9
	Ours	**94.1**	**97.8**	**99.1**	97.2	**95.3**	92.5	**95.7**	**98.1**	**98.8**	**96.7**	**91.5**	**96.1**
BG#1-2	MGAN	48.5	58.5	59.7	58.0	53.7	49.8	54.0	61.3	59.5	55.9	43.1	54.7
	GaitSet	79.9	89.8	91.2	86.7	81.6	76.7	81.0	88.2	90.3	88.5	73.0	84.3
	GaitSlice	85.2	92.2	95.3	94.2	87.8	83.8	87.1	93.1	93.4	91.6	80.9	89.5
	GaitGL	88.5	**95.1**	95.9	94.2	**91.5**	85.4	89.0	95.4	97.4	94.3	86.3	92.1
	Ours	**88.7**	94.6	**96.4**	**94.6**	90.5	**86.0**	**89.3**	**95.5**	**97.8**	**95.0**	**86.9**	**92.3**
CL#1-2	MGAN	3.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
	GaitSet	52.0	66.0	72.8	69.3	63.1	61.2	63.5	66.5	67.5	60.0	45.9	62.5
	GaitSlice	63.9	78.9	82.9	81.7	74.5	70.1	73.4	77.5	77.5	73.7	62.5	74.2
	GaitGL	70.7	83.2	87.1	84.7	78.2	71.3	78.0	83.7	83.6	77.1	63.1	78.3
	Ours	**71.9**	**85.0**	**88.3**	**86.7**	**78.7**	**74.7**	**79.8**	**83.8**	**85.4**	**80.6**	**65.2**	**80.0**

Table 4. Averaged rank-1 accuracies on CASIA-B under ST setting, 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	CNN-LB	54.8	-	-	77.8	-	64.9	-	76.1	-	-	-	68.4
	GaitSet	64.6	83.3	90.4	86.5	80.2	75.5	80.3	86.0	87.1	81.4	59.6	79.5
	GaitSlice	75.7	84.5	92.3	91.3	82.8	77.1	83.1	89.3	91.0	86.2	71.4	84.1
	GaitGL	77.0	**87.8**	93.9	92.7	**83.9**	78.7	**84.7**	91.5	92.5	89.3	74.4	86.0
	Ours	**77.4**	87.6	**94.0**	**92.8**	**83.9**	**78.8**	84.6	**91.7**	**92.6**	**89.5**	**74.8**	**86.2**
BG#1-2	GaitSet	55.8	70.5	76.9	75.5	69.7	63.4	68.0	75.8	76.2	70.7	52.5	68.6
	GaitSlice	67.8	75.0	81.7	82.6	73.8	66.3	73.3	80.6	80.1	75.5	62.1	74.4
	GaitGL	68.1	**81.2**	**87.7**	**84.9**	76.3	70.5	76.1	**84.5**	87.0	83.6	65.0	78.6
	Ours	**68.8**	79.9	86.8	**84.9**	**77.4**	**71.4**	**76.4**	84.2	**87.3**	**84.4**	**67.2**	**78.9**
CL#1-2	GaitSet	29.4	43.1	49.5	48.7	42.3	40.3	44.9	47.4	43.0	35.7	25.6	40.9
	GaitSlice	42.9	55.7	62.2	59.1	54.9	51.3	55.6	55.9	53.6	48.4	35.4	52.3
	GaitGL	46.9	58.7	66.6	65.4	58.3	54.1	59.5	62.7	61.3	57.1	40.6	57.4
	Ours	**47.9**	**60.2**	**68.7**	**68.2**	**61.8**	**58.0**	**63.4**	**66.3**	**63.6**	**59.4**	**43.0**	**60.0**

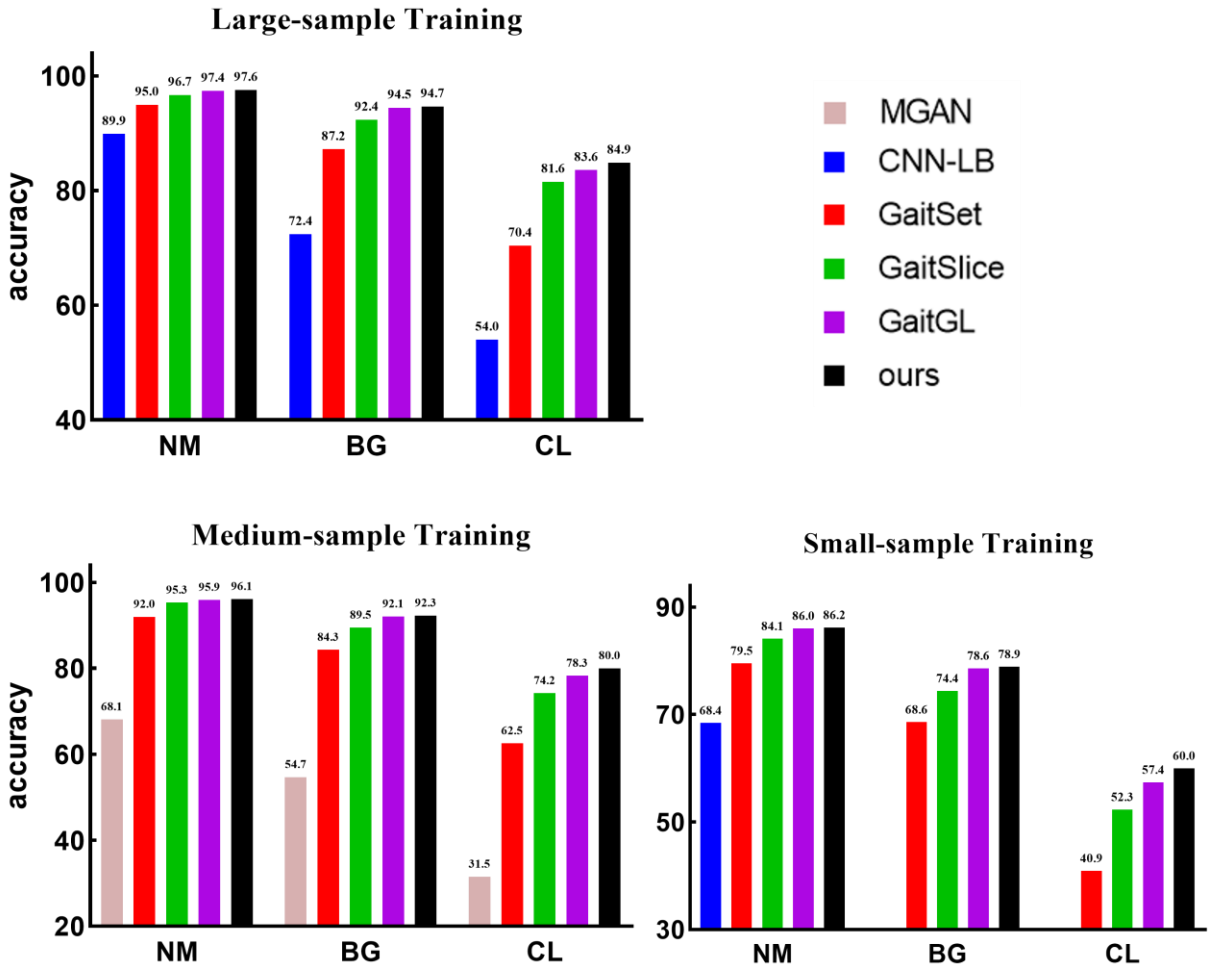

Figure 5. Average rank-1 recognition accuracy of our method compared to other state-of-the-art gait recognition models on CASIA-B under three partition settings.

As can be seen from Table 2, our model obtains very nice results by using LT. Under the three walking conditions NM, BG, and CL, the average recognition accuracy reached 97.6%, 94.7%, and 84.9%, which outperformed GaitGL [5] by 0.2%, 0.3%, and 1.3%, respectively. Comparison results show that the biggest breakthrough of our method is to improve recognition accuracy under CL conditions. Due to the presence of occlusion, the recognition rate of the existing models under CL conditions is low, but in our method, the acquisition of key frames reduces the interference of redundant information, and thus improves the recognition accuracy.

From Tables 3 and 4, it can be seen that under ST and MT settings, compared with existing methods, our model does not achieve an absolute advantage in NM and BG conditions. However, in the CL condition, our model has an average recognition accuracy of 80.0%(MT) and 60.0%(ST), and when compared with the best-performing GaitGL [5], the recognition accuracy is improved by 1.7%(MT) and 2.6%(ST).

Moreover, compared with other views, the existing models have the lowest recognition rate in the view of 0° and 180° due to excessive interference information. In this paper, our method improves the recognition rate in most views, especially at 0° and 180°. For example, compared with GaitGL, the recognition rate in the three walking conditions was improved by 1.2%, 0.4%, and 3.9%, respectively, in the 180° view and LT setting.

In summary, the results of the comparative analysis show that our model outperforms the existing gait recognition models in the LT, MT, and ST settings, especially in complex application scenarios (such as coat-wearing walking conditions and the view of 0° and 180°); this demonstrates the effectiveness and superiority of our model.

4.3.2. Experimental Results on OU-MVLP Dataset

In order to verify the generalization of the proposed method, we further evaluate the performance of our method on the OUMVLP dataset. As shown in Table 5, compared with the existing models (including GEINet [34], GaitSet [2], GaitPart [3], GaitSlice [4], and GaitGL [5]), our model achieves the highest accuracy in various views. Especially in the view with less discriminative information and too much redundant information, such as 0°, 90°, 180°, and 270°. The comparative analysis results on the OU-MVLP dataset show that our model has good generalization.

Table 5. Averaged rank-1 accuracies on OU-MVLP, excluding identical-view cases.

	All 14 Gallery Views														
	0°	**15°**	**30°**	**45°**	**60°**	**75°**	**90°**	**180°**	**195°**	**210°**	**225°**	**240°**	**255°**	**270°**	**Mean**
GEINet	11.4	29.1	41.5	45.5	39.5	41.8	38.9	14.9	33.1	43.2	45.6	39.4	40.5	36.3	35.8
GaitSet	79.5	87.9	89.9	90.2	88.1	88.7	87.8	81.7	86.7	89.0	89.3	87.2	87.8	86.2	87.1
GaitPart	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GaitSlice	84.1	89.0	91.2	91.6	90.6	89.9	89.8	85.7	89.3	90.6	90.7	89.8	89.6	88.5	89.3
GaitGL	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
Ours	**86.3**	**90.8**	**91.3**	**91.6**	**91.2**	**91.0**	**90.7**	**89.4**	**89.5**	**90.5**	**90.6**	**90.0**	**89.8**	**89.3**	**90.1**

4.4. Ablation Experiment

To verify the effectiveness of MTA and MFAM in the proposed framework, several ablation studies with various settings will be conducted on CASIA-B. In MTA, we study the influence of different size convolution on the model performance. In MFAM, we studied the influence of different distance measurement methods and different data normalization methods on the recognition rate. The experimental results and analysis are as follows.

4.4.1. Efficacy of MTA

In order to aggregate context information with different scales and adapt the model to complex human motion patterns, we introduced the Multi-scale Temporal Aggregation module in the proposed method. To analyze the appropriate parameter setting in MTA operation, three controlled experiments are conducted in experiment Group A.

As shown in Table 6, to verify the effectiveness of the MTA module, we design the comparison experiment by implementing methods with different convolutional strategies on the LT setting. Specifically, experiments A-a, b, and c only adopted one convolution kernel with different sizes, and the comparison results with experiment A-d demonstrates the effectiveness of aggregating multi-scale contextual information operations. In addition, the comparison of experiments A-a, b, and c show that the expansion of the convolution size in MTA will lead to a decrease in the recognition accuracy of the model in complex scenes. Therefore, we choose the parameter settings in experiment A-b as the convolution combination.

Table 6. Ablation experiments for Multi-scale Temporal Aggregation. Control condition: the different convolutional combination strategies.

Group A	**kernel_1**	**kernel_2**	**kernel_3**	**NM**	**BG**	**CL**
a	(3,1,1)	-	-	97.5	94.5	84.5
b	-	(5,1,1)	-	97.5	94.6	84.2
c	-	-	(7,1,1)	**97.7**	94.5	84.1
d	(3,1,1)	(5,1,1)	(7,1,1)	97.6	**94.8**	**84.9**

4.4.2. Efficacy of MFAM

Video sequences are rich in semantic information but they also bring too many redundant features. In this paper, we introduced the Metric-based Frame Attention Mechanism

to extract key frame features from sequences. As discussed in Section 3.3, MFAM calculates the importance score of each frame by measuring the Euclidean distance between frame-level features and sequence-level features. To analyze the rationality of the MFAM using Euclidean distance to calculate the importance score, we employ cosine similarity instead of Euclidean distance, as follows:

$$D_t = \frac{p_t \cdot P}{|p_t| \cdot |P|} \tag{6}$$

Moreover, normalization of the data can eliminate the undesirable effects caused by odd sample data. In this section, we introduced the Z-score normalized processing of the numeric distance, the formula is expressed as:

$$D'_t = \frac{D_t - \mu}{\sigma} \tag{7}$$

where μ and σ represent the mean and variance of the numerics, respectively.

As shown in Table 7, to verify the effectiveness of the MFAM module, we design the comparison experiment by implementing methods with different importance score calculation strategies on the LT setting.

Table 7. Ablation experiments for Metric-based Frame Attention Mechanism. Control condition: the different data normalization methods and different distance metric strategies. D_E and D_C: represent the Euclidean distance and cosine similarity, respectively. N_m and N_z: represent the max-min normalized and Z-score normalized, respectively.

Distance	Normalization	NM	BG	CL
D_E	N_m	97.6	**94.8**	**84.9**
	N_z	**97.7**	94.6	84.3
D_C	N_m	97.5	94.7	84.2
	N_z	97.6	94.6	84.2

On the one hand, the comparison shows that employing Euclidean distance to measure the similarity between features is conducive to obtaining better performance of the model. This is because the model employs Euclidean distance as the accuracy evaluation standard. On the other hand, the comparison results of different data normalization methods show that a Z-score normalized is only applicable to NM scenarios. In complex application scenarios such as BG and CL, max-min normalized shows superior recognition accuracy. Therefore, we chose the combination of Euclidean distance and max-min normalized to calculate the importance score of each frame.

4.5. Practicality Experiments

In real-world settings, it may be difficult to acquire a sufficient number of frames for gait recognition. To verify the practicability of the proposed model, we select a certain number of frames for each subject during the testing phase.

As shown in Table 8, we selected different numbers of frames as input. The results show that the model achieves accuracies of 57.4%, 50.4%, and 33.2% in the three walking conditions when inputting 10 frames. This is because the model uses an encoder composed of 3D convolutions. When the input frame is insufficient, it is difficult for the model to extract gait temporal features. When the input reaches 30 frames, the model recognition accuracies also achieve 94.5%, 90.1%, and 76.6%. When the input exceeds 70 frames, the model recognition accuracy also tends to be stable.

Table 8. The result when the number of input frames is different. Results are rank-1 accuracies (%) averaged on all 11 views, excluding identical-view cases.

Number of Frames	NM	BG	CL
10	57.4	50.4	33.2
20	86.3	80.1	60.8
30	94.5	90.1	76.6
50	96.9	93.4	82.5
70	97.4	94.3	83.5
100	97.5	94.5	83.9

4.6. Portability Experiments

It is worth noting that our MFAM operation can be used in some state-of-the-art gait recognition models [2–5]. In the GaitSlice model proposed by Li et al., the RFAM module is designed to calculate the importance score of each frame by the frame attention network. To compare the performance differences between the two attention mechanisms, three controlled experiments are conducted. In the experiments, we use the experimental results of the GaitSlice [4] model (denoted as GaitSlice* and achieves accuracies of 96.6%, 91.7%, and 80.7%, respectively, under three walking conditions) with the RFAM module removed as the baseline.

As shown in Figure 6, under the action of a single RFAM module, the recognition rates of the three walking scenarios are increased by 0.06%, 0.71%, and 0.86%, respectively. Under the action of our MFAM module, the recognition rates are increased by 0.16%, 0.75%, and 1.13%, which outperform RFAM by 0.1%, 0.04%, and 0.27%, respectively. Under the joint action of the two attention modules, the recognition rate is increased by 0.25%, 0.81%, and 1.28%, respectively. The above results demonstrate the effectiveness and superiority of our proposed MFAM operation.

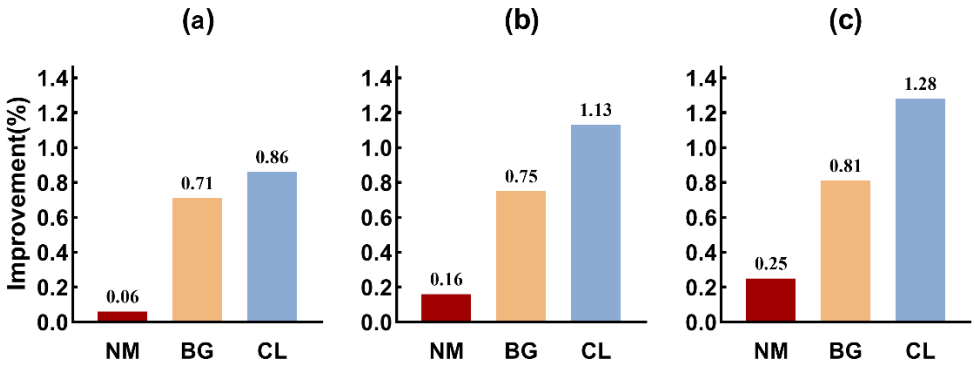

Figure 6. Performance comparison experiment between MFAM and RFAM. Results are improvement compared to baseline (GaitSlice*) of rank-1 accuracies without view variation, excluding the identical-view cases. (**a**) Accuracy improvement of introducing RFAM (**b**); Accuracy improvement of introducing MFAM; (**c**) Accuracy improvement of introducing MFAM and RFAM.

4.7. Visualization

In order to better understand the role of the MFAM module, the importance scores of several frames are given in Figure 7. It is worth noting that we horizontally divide the human body into 32 parts in the model, then calculate each part's importance score. For the convenience of illustration, the average of the importance scores of the adjacent eight parts is calculated in Figure 7. In other words, we show the weights of the four parts of the human body. It can be seen from Figure 7 that the importance scores of each frame in the sequence are different, indicating that the frames contain different degrees of semantic information.

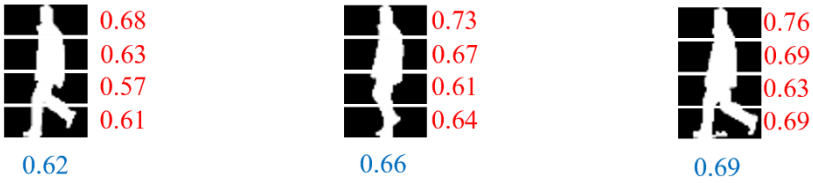

Figure 7. Importance scores for frames. The red words are the importance score of each part, and the blue words are the importance score of each frame.

5. Conclusions

In this paper, we propose a new gait recognition framework, which models gait temporal features and highlights the key frames in the sequence to improve recognition accuracy. Specifically, the proposed MTA module extracts multi-scale context information through parallel convolution and carries out information exchange and fusion, so that the model can adapt to the changes of complex motion and realistic factors. In MFAM, the importance score of each frame is calculated by the distance between frame-level features and sequence-level features. After that, more discriminative gait features are extracted by reweighting key frames for each frame. Finally, experiments are conducted on the widely adopted public databases, CASIA-B and OUMVLP, which experimentally demonstrate the superiority of the proposed method. Nevertheless, the CASIA-B and OUMVLP datasets were collected in an indoor environment. Although they include different walking conditions such as viewing angles, clothing, and carrying objects, there are still differences with the pedestrian data under real conditions. In future work, the model will be optimized for the dataset collected in the open environment.

Author Contributions: Conceptualization, T.W. and R.C.; methodology, T.W. and H.Z.; software, R.L. and R.C.; validation, J.Z. and H.L.; formal analysis, R.L. and J.W.; investigation, J.Z. and J.W.; resources, R.L. and H.Z.; data curation, T.W. and J.Z.; writing—original draft preparation, T.W. and H.L.; writing—review and editing, T.W. and H.Z.; supervision, H.Z.; project administration, T.W. and H.Z.; funding acquisition, H.Z. and R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No.62072122), the Scientific and Technological Planning Projects of Guangdong Province (2021A0505030074), the Scientific Research Capability Improvement Project of Guangdong Key Construction subject (2021ZDJS025), the Postgraduate Education Innovation Plan Project of Guangdong Province (2020SFKC054), the Special Projects in Key Fields of Ordinary Universities of Guangdong Province under Grant (2021ZDZX1087) and the Special Projects in key Fields of Department of Education of Guangdong Province (2022ZDZX1013).

Data Availability Statement: The data are available upon request from the corresponding author.

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

References

1. Zhang, Z.; Tran, L.; Liu, F.; Liu, X. On learning disentangled representations for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 345–360. [CrossRef] [PubMed]
2. 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 27–1 February 2019; pp. 8126–8133.
3. Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; He, Z. GaitPart: Temporal part-based model for gait recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 14225–14233.
4. Li, H.; Qiu, Y.; Zhao, H.; Zhan, J.; Chen, R.; Wei, T.; Huang, Z. GaitSlice: A gait recognition model based on spatio-temporal slice features. *Pattern Recognit.* **2022**, *124*, 108453. [CrossRef]
5. Lin, B.; Zhang, S.; Yu, X. Gait recognition via effective global-local feature representation and local temporal aggregation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 14648–14656.

6. Wang, K.; Wang, S.; Zhang, P.; Zhou, Z.; Zhu, Z.; Wang, X.; Peng, X.; Sun, B.; Li, H.; You, Y. An efficient training approach for very large scale face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4083–4092.
7. He, M.; Zhang, J.; Shan, S.; Chen, X. Enhancing Face Recognition with Self-Supervised 3D Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 4062–4071.
8. Öztürk, H.İ.; Selbes, B.; Artan, Y. MinNet: Minutia Patch Embedding Network for Automated Latent Fingerprint Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1627–1635.
9. Chen, S.; Guo, Z.; Li, X.; Yang, D. Query2Set: Single-to-Multiple Partial Fingerprint Recognition Based on Attention Mechanism. *IEEE Trans. Inf. Secur.* **2022**, *17*, 1243–1253. [CrossRef]
10. Han, J.; Bhanu, B. Individual recognition using gait energy image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *28*, 316–322. [CrossRef] [PubMed]
11. Bashir, K.; Xiang, T.; Gong, S. Gait recognition using gait entropy image. In Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), London, UK, 3 December 2009.
12. Wang, C.; Zhang, J.; Wang, L.; Pu, J.; Yuan, X. Human identification using temporal information preserving gait template. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 2164–2176. [CrossRef] [PubMed]
13. 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 (ICPR'06), Hong Kong, China, 20–24 August 2006; Volume 4, pp. 441–444.
14. Takemura, N.; Makihara, Y.; Muramatsu, D.; Echigo, T.; Yagi, Y. Multi-view large population gait dataset and its performance evaluation for crossview gait recognition. *IPSJ Trans. Comput. Vis. Appl.* **2018**, *10*, 4. [CrossRef]
15. Rong, Z.; Vogler, C.; Metaxas, D. Human Gait Recognition. In Proceedings of the Conference on Computer Vision & Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004.
16. Hong, C.; Yu, J.; Tao, D.; Wang, M. Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval. *IEEE Trans. Ind. Electron.* **2014**, *62*, 3742–3751.
17. Huang, Z.; Xue, D.; Shen, X.; Tian, X.; Li, H.; Huang, J.; Hua, X.-S. 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; pp. 14920–14929.
18. Hou, S.; Cao, C.; Liu, X.; Huang, Y. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *Lecture Notes in Computer Science*; Springer: Cham, Switzerland, 2020; pp. 382–398.
19. Zhang, Z.; Tran, L.; Yin, X.; Atoum, Y.O.; Wan, J.; Wang, N.; Liu, X. Gait recognition via disentangled representation learning. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
20. 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]
21. Liao, R.; Cao, C.; Garcia, E.B.; Yu, S.; Huang, Y. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2017.
22. 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]
23. 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; pp. 7291–7299.
24. Guler, R.A.; Neverova, N.; Kokkinos, I. 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; pp. 7297–7306.
25. 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 (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 4165–4169.
26. Masood, H.; Farooq, H. Utilizing Spatio Temporal Gait Pattern and Quadratic SVM for Gait Recognition. *Electronics* **2022**, *11*, 2386. [CrossRef]
27. Song, G.; Leng, B.; Liu, Y.; Hetang, C.; Cai, S. Region-based Quality Estimation Network for Large-scale Person Re-identification. *arXiv preprint* **2017**, arXiv:1711.08766. [CrossRef]
28. Ding, Y.; Hou, S.; Yang, X.; Du, W.; Wang, C.; Yin, G. Key Frame Extraction Based on Frame Difference and Cluster for Person Re-identification. In Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), Atlanta, GA, USA, 18–21 October 2021; pp. 573–578.
29. Wang, X.; Feng, S.; Yan, W.Q. Human Gait Recognition Based on Self-Adaptive Hidden Markov Model. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2021**, *18*, 963–972. [CrossRef] [PubMed]
30. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
31. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Wu, Z.; Huang, Y.; Wang, L.; Wang, X.; Tan, T. A comprehensive study on cross-view gait based human identification with deep CNNs. *IEEE TPAMI* **2017**, *39*, 209–226. [CrossRef] [PubMed]

33. He, Y.; Zhang, J.; Shan, H.; Wang, L. Multi-task GANs for view-specific feature learning in gait recognition. *IEEE TIFS* **2019**, *14*, 102–113. [CrossRef]
34. 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 (ICB), Halmstad, Sweden, 13–16 June 2016; pp. 1–8.