Context-Sensitive Temporal Feature Learning for Gait Recognition

Xiaohu Huang*[1], Duowang Zhu*[1], Hao Wang[1], Xinggang Wang[1], Bo Yang[2], Botao He[2], Wenyu Liu[1], and Bin Feng†[1]

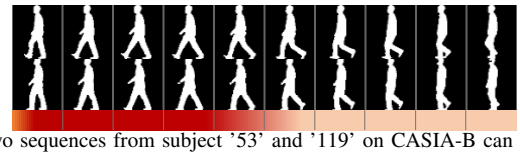
[1]*Huazhong University of Science and Technology*
[2]*Wuhan FiberHome Digital Technology Co., Ltd*

Abstract

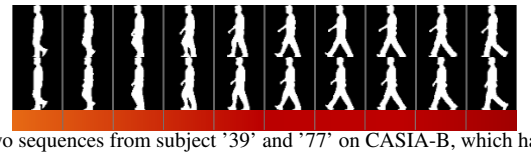
*Although gait recognition has drawn increasing research attention recently, it remains challenging to learn discriminative temporal representation since the silhouette differences are quite subtle in spatial domain. Inspired by the observation that humans can distinguish gaits of different subjects by adaptively focusing on temporal sequences with different time scales, we propose a context-sensitive temporal feature learning (**CSTL**) network in this paper, which aggregates temporal features in three scales to obtain motion representation according to the temporal contextual information. Specifically, CSTL introduces relation modeling among multi-scale features to evaluate feature importances, based on which network adaptively enhances more important scale and suppresses less important scale. Besides that, we propose a salient spatial feature learning (**SSFL**) module to tackle the misalignment problem caused by temporal operation,* e.g.*, temporal convolution. SSFL recombines a frame of salient spatial features by extracting the most discriminative parts across the whole sequence. In this way, we achieve adaptive temporal learning and salient spatial mining simultaneously. Extensive experiments conducted on two datasets demonstrate the state-of-the-art performance. On CASIA-B dataset, we achieve rank-1 accuracies of 98.0%, 95.4% and 87.0% under normal walking, bag-carrying and coat-wearing conditions. On OU-MVLP dataset, we achieve rank-1 accuracy of 90.2%. The source code will be published at* https://github.com/OliverHxh/CSTL.

1. Introduction

Gait recognition is a long-distance biological identification technology, which relies on the walking patterns of human beings, and reveals great application potential on identity recognition [20, 1, 22]. Although gait recognition has drawn increasing research attention recently, it remains challenging to learn discriminative temporal representation since the silhouette differences in spatial domain are quite subtle.



(a) Two sequences from subject '53' and '119' on CASIA-B can be distinguished relying on short-term temporal clues, e.g., several frames at the beginning.



(b) Two sequences from subject '39' and '77' on CASIA-B, which have to be distinguished relying on long-term temporal clues, e.g., all of the frames.

Figure 1. Illustration that humans can distinguish gaits of various subjects by adaptively focusing on temporal fragments with different time scales. Color bar indicates the human focus distribution. Darker color represents more attention needed for corresponding frames. Best viewed in color.

Moreover, as mentioned in [6], body parts possess diverse motion patterns which requires temporal modeling to take multi-scale representation into consideration. Multi-layer temporal convolution has been widely used in current methods [6, 28, 18, 31, 32] to model temporal information in multiple scales. They aggregated multi-scale temporal features in a summation or a concatenation way. However, these manners are not flexible enough to adapt to variation of complex motion and realistic factors, *i.e.*, occlusion of clothing and change of camera viewpoints, since the fusion method of multi-scale features is fixed. Thus, the performance is hindered especially considering gait is a kind of fine-grained motion pattern, whose identification of sub-

*Contributed equally.
†Corresponding author: fengbin@hust.edu.cn.

jects depends on the diverse expression on tiny motion of local body.

It can be seen from life experience that humans distinguish gait sequences of different subjects by adaptively focusing on temporal fragments with different time scales. A qualitative illustration is given in Fig.1, where voting results from seven volunteers are used to calculate the focus distribution. In Fig.1(a), the differences between two gait sequences are so obvious that we can distinguish them by observing several frames from beginning. On the contrary, in Fig.1(b), differences between two sequences are quite subtle that we have to observe more frames to distinguish them. Therefore, in this situation, short-term clues are not enough to make a distinction between the two subjects. Long-term features need to be considered since they provide richer temporal information. Hence, the adaptive adjustment among multi-scale temporal features leads to flexible focus along temporal dimension, which offers a new perspective for gait modeling.

Motivated by such observation, we propose a context-sensitive temporal feature learning (**CSTL**) network for gait recognition. The core idea of this method is to integrate multi-scale temporal features according to the contextual information along temporal dimension, which allows information communications among different scales. Here, contextual information is obtained by evaluating the relations among multi-scale temporal features, which reflects diverse motion information existing in context features. CSTL produces temporal features in three temporal scales, *i.e.*, frame-level, short-term and long-term, which are complementary to each other. The frame-level features retain frame characteristics at each time instant. The short-term features capture local temporal contextual clues, which are sensitive to temporal locations and beneficial to model micro motion patterns. The long-term features, on behalf of motion features across all frames, reveal global action periodicities of different body parts, which are invariant for temporal locations. Then, the relation modeling among these temporal features guides the network to adaptively enhance or suppress temporal features with different scales, then generates appropriate temporal descriptions for motion learning on different body parts. This method provides the possibility of modeling complex motion, which makes it very suitable for gait recognition.

Further, during the investigation of temporal modeling, we notice the misalignment problem in temporal modeling that has not been investigated in gait recognition yet. As shown in Fig.2, the same pixel locations from different frames may correspond to different foregrounds and backgrounds. Naturally, the utilization of temporal operations, *e.g.*, temporal convolutions and temporal poolings, may result in blurry and overlapped appearances. To address such issue, we propose a salient spatial feature learning (**SSFL**) module to select discriminative spatial clues across the whole sequence, which is considered as a supplement to remedy the corruption in appearance features.

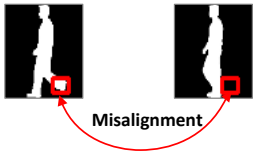


Figure 2. Illustration of misalignment problem caused by temporal convolution, since pixels of same spatial locations in different frames may correspond to different semantic content.

The adaptive temporal modeling and salient spatial learning provide complementary properties for each other. On one hand, CSTL mainly considers temporal modeling and SSFL focuses on spatial learning. Specifically, CSTL produces temporal aggregation of multi-scale clues which describes motion patterns, and SSFL generates recombinant frame features which involve with still images. On the other hand, CSTL aggregates temporal clues in a soft-attention way and SSFL selects salient spatial features in a hard-attention manner. In a word, by jointly investigating motion learning and spatial mining simultaneously, we achieve outstanding performance over the existing methods.

The major contributions of this paper can be summarized as the following three aspects:

- In this paper, we propose a temporal modeling network CSTL to fuse multi-scale temporal features in an adaptive way, which considers the cross-scale contextual information as a guidance for temporal aggregation.
- we propose a salient spatial feature learning (SSFL) module to remedy the misalignment problem caused by temporal operation. SSFL extracts salient spatial features from different frames to form a recombinant frame which maintains high quality spatial features.
- Extensive experiments conducted on two popular datasets CASIA-B [30] and OU-MVLP [24] demonstrate the state-of-the-art performance of our method. And further ablation experiments prove the effectiveness of the proposed modules.

2. Related Work

Gait Recognition. Current gait recognition methods can be categorized into two types: model-based and appearance-based. **Model-based** models [17, 16, 25] were proposed to model walking patterns and body structures of humans based on extracted pose information [2, 23, 3]. Model-based methods are robust to variations of clothing and camera viewpoints. However, due to the inaccurate key point estimation results from low-quality images and the missing

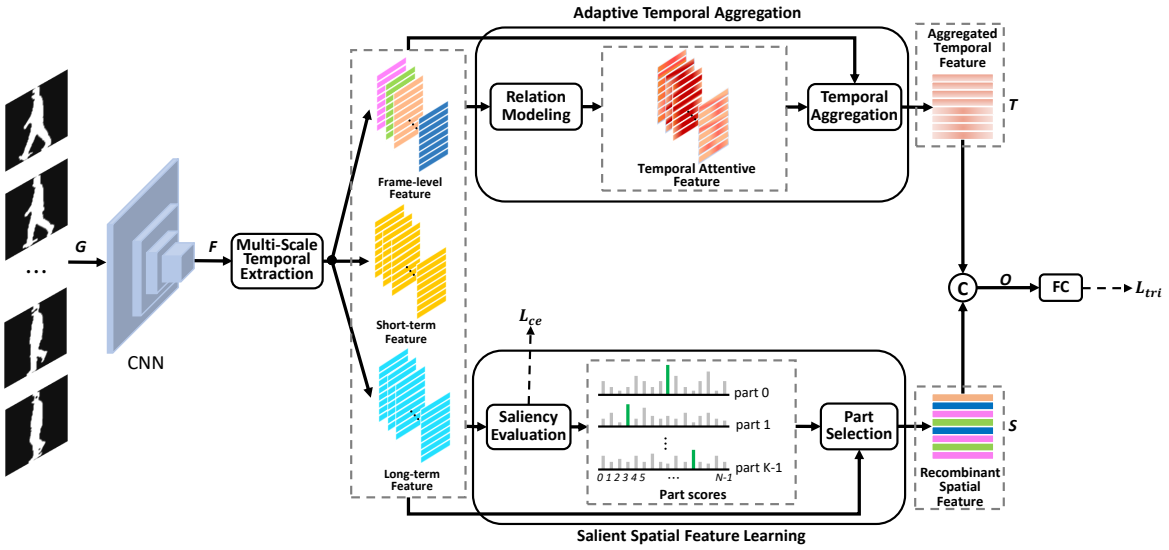


Figure 3. Overview of CSTL. Arrows, G, P, T, S and O denote operations, input gait sequence, pooled part-level features, temporal aggregated features and recombinant spatial features respectively. L_{tri} and L_{ce} represent triplet loss and cross-entropy loss respectively.

of identity-related shape information, model-based methods are usually inferior to appearance-based methods in performance comparison. **Appearance-based** models [4, 6, 12, 18, 31, 27, 9, 10, 29, 13] extracted spatio-temporal features based on RGB images or binary silhouettes by CNN networks or handcrafted algorithms. [9, 10, 29, 13, 15] generated Gait Energy Image (GEI) [9] by temporal average pooling, which greatly reduced the computation cost but lost discriminative expression. [4, 6, 12, 18, 31] processed gait sequences frame by frame, which maintained the frame-level discriminative feature in a large extent . Our approach belongs to appearance-based method and takes silhouette sequences as input.

Temporal Modeling. Current literatures proposed different strategies for gait temporal modeling, including 1D convolutions, LSTMs and 3D convolutions. GaitSet and GLN [4, 12] considered a gait sequence as a unordered set, which mainly focused on spatial modeling but neglected inter-frame dependency modeling. GaitPart [6] and Wu *et al.* [28] extracted local temporal clues by 1D convolutions and aggregated them in a summation or a concatenation manner. LSTM networks were applied in [31, 32] to achieve long-short temporal modeling, which fused temporal clues by temporal accumulation. With the help of stacked 3D blocks, MT3D [18] incorporated temporal information with small and large scales, then concatenated these features as outputs. In summary, there are obvious shortcomings in learning flexible and robust multi-scale temporal features of current methods, which were incapable of satisfying temporal modeling requirements for gait motion.

Compared to the above methods, in the paper, CSFL utilizes temporal features in three scales: frame-level, short-term and long-term. Such rich temporal clues enable our network to obtain diverse motion learning capability. And by employing cross-scale relation modeling of multi-scale temporal clues, we adjust the feature expression to emphasize different frames along temporal dimension, then produce appropriate sequence-level motion representation in a weighted summation way.

Spatial Preserving. A problem related to temporal modeling is spatial misalignment, which may degrade performance severely in person related recognition task, e.g. Person Re-identification. In video-based Person Re-identification, different methods were proposed to maintain the clearness of spatial features. In AP3D [8], researchers proposed Appearance Preserving Module (APM) to mitigate the misalignment problem in temporal modeling. APM used a feature similarity calculation strategy to match the foregrounds in continuous frames within a local window based on the color, texture and illumination *et al.* Chen *et al.* [5] proposed a method dubbed Adversarial Feature Augmentation (AFA) to capture motion coherence by a adversarial form.

Different from these strategies, in our approach, SSFL selects discriminative spatial local features to maintain the spatial characteristics of subjects, which is feasible for binary inputs. And this operation is parallel to temporal modeling process, thus would not affect temporal feature extraction.

3. Method

In this section, we firstly describe the overall pipeline of our method, then illustrate the detailed structure of each

component in the network.

3.1. Network Pipeline

The overview structure of our method is presented in Fig.3. A batch of B gait samples of N frames are fed into the network as input, which is denoted as $G \in \mathbb{R}^{B\times N\times H\times W}$. H and W denote the height and width of each input frame respectively. Firstly, G is passed through a 2D CNN with 4 layers to produce feature $F \in \mathbb{R}^{B\times N\times C\times H/2\times W/2}$, where C denotes the number of feature channels. Afterwards, we implement a multi-scale temporal extraction module on F to generate temporal features with three different temporal scales, *i.e.*, frame-level, short-term and long-term, which are denoted as T_f, T_s and T_l respectively. T_f, T_s and T_l all own size of $\mathbb{R}^{B\times N\times C\times K}$, where K denotes the number of horizontal division feature parts that correspond to body parts in some extent. Next, temporal features are taken as the inputs for Adaptive Temporal Aggregation (ATA) and Salient Spatial Feature Learning (SSFL) blocks through which we obtain temporal aggregated feature $T \in \mathbb{R}^{B\times C\times K}$ and recombinant spatial salient feature $S \in \mathbb{R}^{B\times C\times K}$ correspondingly. Temporal aggregated feature T is a weighted summarization of whole sequence features by the importance of each feature map to represent the discriminative information in temporal domain. Spatial salient feature S is recombined by selecting the most salient spatial parts which maintain rich undistorted silhouette information. Finally, S and T are concatenated along channel dimension as outputs O.

3.2. Multi-Scale Temporal Extraction

As discussed in Sec.3.1, we aim to enrich the diversity of temporal features. Firstly, we divide F into K parts, then apply Global Max Pooling (GMP) and Global Average Pooling (GAP) to obtain part-level pooling features $P \in \mathbb{R}^{B\times N\times C\times K}$, where P_b^n represents part-level features of the n-th frame in the b-th sample. As shown in Fig.4, the frame-level features are the duplicate of P, which do not get involved with temporal operation, thus the appearance characteristics of each time instant are well-maintained.

In order to capture short-term temporal features, we apply two serial 1D convolutions with kernel size of 3, and add the features after each 1D convolution as T_s. Obtaining short-term features enables the network to focus on short period temporal motion patterns and subtle changes with perceptive fields of 3 and 5.

The long-term feature extraction is based on the combination of all frames. Firstly, a Multi-layer Perceptron (MLP) followed by a Sigmoid function is applied on P to evaluate the importance of different frames. Next, the weighted summation of all frames by the importance scores is utilized as the long-term temporal features T_l, which is

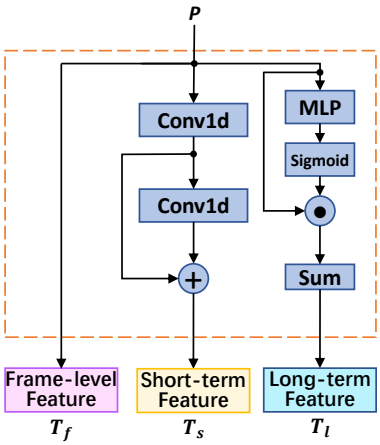


Figure 4. Details of Multi-scale Temporal Feature Learning. The detailed structure of producing temporal features in three levels.

formulated as:

$$T_l^b = \frac{\sum_{n=1}^{N} Sigmoid(MLP(P_b^n)) \odot P_b^n}{\sum_{n=1}^{N} Sigmoid(MLP(P_b^n))}, \tag{1}$$

where $\odot$ denotes dot product. It should be noted that, T_l^b is invariant for all frames in the b-th sample, which describes global motion cues. After that, we obtain temporal features of three levels, *e.g.*, T_f, T_s and T_l, for subsequent ATA and SSFL blocks.

3.3. Adaptive Temporal Aggregation

Relation Modeling. In this part, we utilize multi-scale temporal features to explore feature relations, which enable information exchanging among different temporal scales. As discussed in [6], different body parts own various motion patterns, which indicates the diverse expressions are needed for temporal modeling. Intuitively, feature relation modeling provides a variety of temporal perceptive fields. Therefore, the interaction of different type of features would effectively enrich the diversity of temporal representation, thus produce suitable motion expression for human body.

As shown in Fig.5, the cross-scale relation modeling produces individual scores for evaluating importance of temporal features from different scales. Such relation modeling leverages rich temporal information in an efficient way, which involves with diverse temporal granularities for describing motion patterns of different body parts adaptively. Firstly, we apply information flowing among temporal features from top to bottom:

$$\begin{aligned} \widetilde{T}_f &= T_f \\ \widetilde{T}_s &= T_f + T_s \\ \widetilde{T}_l &= T_f + T_s + T_l. \end{aligned} \tag{2}$$

Then, we learn temporal importance weight for each temporal scale by considering the contextual information of the

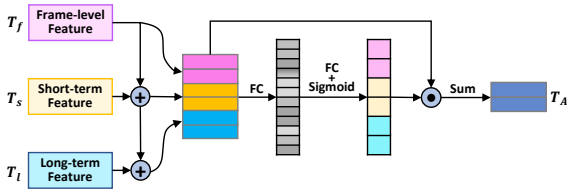


Figure 5. Structure of relation modeling across three temporal scales.

three temporal scales, which is implemented with two fully connected layers and a Sigmoid function:

$$W_T = Sigmoid(FC(FC(\widetilde{T}_f \copyright \widetilde{T}_s \copyright \widetilde{T}_l))), \tag{3}$$

where $W_T \in \mathbb{R}^{B\times N\times 3\times C\times K}$ and $W_T^{b,n}$ denote the temporal importance weights of the n-th frame in the b-th sample. W_T incorporates importance weights of the three temporal scales, which is denoted as $W_{T,1}$, $W_{T,2}$ and $W_{T,3}$ respectively. Afterwards, we obtain attentive temporal features by a soft-attention manner:

$$T_A^{b,n} = \widetilde{T}_f^{b,n} \odot W_{T,1}^{b,n} + \widetilde{T}_s^{b,n} \odot W_{T,2}^{b,n} + \widetilde{T}_l^{b,n} \odot W_{T,3}^{b,n}. \tag{4}$$

Based on the cross-scale temporal aggregation, we obtain sequence-level representation in a weighted summation manner for the b-th sample:

$$T_b = \frac{\sum_{n=1}^{N} T_A^{b,n}}{\sum_{n=1}^{N}\sum_{i=1}^{3} W_{T,i}^{b,n}}, \tag{5}$$

where $T = \{T_b | b = 1, ...B\}$ and $T \in \mathbb{R}^{B\times C\times K}$. The temporal relation modeling encourages our network to generate motion features with adaptive temporal perceptive fields, thus features are highlighted or suppressed adaptively for motion learning.

3.4. Salient Spatial Feature Learning

In this section, we aim to extract salient spatial parts to mitigate the damage in appearance features.

Discussion. Intuitively, in order to remedy the corrupted spatial features, we should select an individual frame as the methods in [7, 14]. However, due to the camera viewpoint and motion occlusion, *e.g.* occlusion of arms, legs and torso, a single frame is probably incapable of expressing appearance features for all body parts clearly. Actually, the high quality body parts appear and disappear from frame to frame. Therefore, by utilizing such inherent motion characteristics, we select salient body parts across the whole sequence to recombine a frame of discriminative features instead of directly selecting one frame.

Operation. Temporal clues provide contextual information for evaluating the discrimination of each frame. Therefore, we apply MLP with Sigmoid function on the temporal features of the three levels for producing part scores of each frame, which is defined as:

$$\begin{aligned} P_s^{b,n} &= Sigmoid(MLP(T_f^{b,n} \copyright T_s^{b,n} \copyright T_l^{b,n})) \\ \widetilde{P}_s^{b,n} &= \frac{P_s^{b,n}}{\sum_{n=1}^{N} P_s^{b,n}}, \end{aligned} \tag{6}$$

where $\widetilde{P}_s^{b,n} \in \mathbb{R}^{1\times K}$ denotes the part scores of the n-th frame in the b-th sample and $\widetilde{P}_s^{b,n,k}$ denotes the k-th part score of the n-th frame on the b-th sample. The values of part scores represent the importance of local parts, thus higher scores indicate clearer spatial representation. In order to supervise the correctness of saliency description, we enforce a fully-connected layer with a cross-entropy loss on the weighted summation of T_f and $\widetilde{P}_s$. Firstly, the weighted part features of the b-th sample with a fully-connected layer is presented as:

$$P_w^b = FC(\sum_{n=1}^{N} T_f^{b,n} \odot \widetilde{P}_s^{b,n}), \tag{7}$$

where $P_w^b \in \mathbb{R}^{C_t\times K}$, and C_t denotes the number of training subjects. Then, cross-entropy loss is applied on P_w^b to produce L_{ce}:

$$L_{ce} = -\sum_{b=1}^{B}\sum_{c=1}^{C_t} y_{b,c} \log(SoftMax(P_w^b))_c, \tag{8}$$

where $y_{b,c}$ indicates the identity information of the b-th sample, which equals 0 or 1.

Afterwards, we obtain part indexes of the highest scores along temporal dimension:

$$x_b^k = \arg\max_n P_s^{b,n,k}, \tag{9}$$

where x_b^k denotes the temporal index of the selected k-th part in the b-th sample. Then, we obtain the recombinant frame feature S_b by the guidance of $\{x_b^k | k = 1, 2, ..., K\}$ in a hard-attention way:

$$S_b = T_f^{b,x_b^1,1} \copyright T_f^{b,x_b^2,2} \cdots \copyright T_f^{b,x_b^K,K}, \tag{10}$$

where © denotes concatenation. Thus, we get recombinant spatial features $S = \{S_b | b = 1, 2, ..., B\}$. S offers supplementary spatial clues for temporal aggregated features T. Triplet loss [11] is employed on the combination of S and T as metric learning loss function. The overall loss function is presented as following:

$$L = L_{ce} + L_{tri} \tag{11}$$

Table 1. Averaged rank-1 accuracies (%) on CASIA-B, excluding identical-view cases.

Gallery NM		Resolution	0 – 180°											Mean
Probe		–	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM	GaitSet[4]	64 × 44	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
		128 × 88	91.4	98.5	98.8	97.2	94.8	92.9	95.4	97.9	98.8	96.5	89.1	95.6
	GaitPart [6]	64 × 44	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 [18]	64 × 44	95.7	98.2	99.0	97.5	95.1	93.9	96.1	98.6	99.2	98.2	92.0	96.7
	GLN [12]	128 × 88	93.2	99.3	**99.5**	**98.7**	96.1	95.6	97.2	98.1	99.3	98.6	90.1	96.9
	CSTL (ours)	64 × 44	97.2	99.0	99.2	98.1	96.2	95.5	**97.7**	98.7	99.2	98.9	96.5	97.8
		128 × 88	**97.8**	**99.4**	99.2	98.4	**97.3**	**95.2**	96.7	**98.9**	**99.4**	**99.3**	**96.7**	**98.0**
BG	GaitSet [4]	64 × 44	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
		128 × 88	89.0	95.3	95.6	94.0	89.7	86.7	89.7	94.3	95.4	92.7	84.4	91.5
	GaitPart [6]	64 × 44	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 [18]	64 × 44	91.0	95.4	97.5	94.2	92.3	86.9	91.2	95.6	97.3	96.4	86.6	93.0
	GLN [12]	128 × 88	91.1	**97.7**	97.8	95.2	92.5	**91.2**	92.4	96.0	97.5	95.0	88.1	94.0
	CSTL (ours)	64 × 44	91.7	96.5	97.0	95.4	90.9	88.0	91.5	95.8	97.0	95.5	90.3	93.6
		128 × 88	**95.0**	96.8	**97.9**	**96.0**	**94.0**	90.5	**92.5**	**96.8**	**97.9**	**99.0**	**94.3**	**95.4**
CL	GaitSet [4]	64 × 44	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
		128 × 88	66.3	79.4	84.5	80.7	74.6	73.2	74.1	80.3	79.7	72.3	62.9	75.3
	GaitPart [6]	64 × 44	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 [18]	64 × 44	76.0	87.6	89.8	85.0	81.2	75.7	81.0	84.5	85.4	82.2	68.1	81.5
	GLN [12]	128 × 88	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	CSTL (ours)	64 × 44	78.1	89.4	91.6	86.6	82.1	79.9	81.8	86.3	88.7	86.6	75.3	84.2
		128 × 88	**84.1**	**92.1**	**91.8**	**87.2**	**84.4**	**81.5**	**84.5**	**88.4**	**91.6**	**91.2**	**79.9**	**87.0**

4. Experiments

4.1. Datasets and Evaluation Metrics.

We conduct experiments on two standard datasets, *i.e.*, CASIA-B [30] and OU-MVLP [24], to verify the superiority of our method. Further ablation experiments are conducted on CASIA-B to demonstrate the positive impact of each component in our method.

CASIA-B. CASIA-B [30] is composed of 124 subjects, and each subject contains 110 sequences with 11 different camera views. Under each camera view, each subject contains three walking conditions, *i.e.*, normal (NM) (6 sequences), walking with bag (BG) (2 sequences) and walking with coat (CL) (2 sequences). During training and testing stage, we follow the protocols in [29]. The samples from the first 74 subjects are considered as train set, and the remaining 50 subjects are considered as test set. At testing phase, the first 4 sequences in NM condition of each subject are regarded as gallery set and the remaining 6 sequences of each subject are used as probe set, including 2 sequences of NM, 2 sequences of BG and 2 sequences of CL.

OU-MVLP. OU-MVLP [24] is composed of 10307 subjects. Each subject contains 28 sequences with 14 camera views, thus each subject contains 2 sequences (index '01' and '02') for each view. The first 5153 subjects are used for training, while the remaining 5154 subjects are for testing. In particular, the sequences with index '01' are regarded as gallery and the sequences with index '02' are regarded as probe set at testing phase.

4.2. Implementation Details

Hyper-parameters. 1) We set the value of B (number of training samples in one iteration) as 64 and 256 on CASIA-B [30] and OU-MVLP [24] datasets respectively. 2) The value of N (input frame number) and K (part division number) are set as 30 and 32. And ablation experiments on K are appended in *supplementary material*. 3) The number of output channels for FCs shown in Fig.3 is set to 256 and 512 respectively for CASIA-B [30] and OU-MVLP [24] datasets. 4) All MLPs follow: FC(c,c/16)->ReLU()->FC(c/16,c). The two FCs in ATA are FC(c,c/16) and FC(c/16,c).

Training Details. **1**) Each frame is aligned as [24] does, and we resize each frame to the size of 64×44 or 128×88. For each input sequence, we follow the frame sampling strategy as [6] does. **2**) We apply separate Batch All (BA_+) triplet loss to train our network. The batch size for training is noted as (p, k), where p denotes the number of sampled subjects and k denotes the number of sampled sequences for each subject. Particularly, (p, k) are set to $(8, 8)$ on CASIA-B and $(32, 8)$ on OU-MVLP. **3**) Since the data amount of OU-MVLP [24] is 20 times larger than that of CASIA-B [30], the numbers of output channels of each layer in 4-layer CNN are set to 32/64, 64/128, 128/256, 128/256 on CASIA-B [30] and OU-MVLP [24] datasets respectively, which follows the design in GaitSet [4] and GLN [6]. And a max pooling layer with stride of 2 is appended after the second convolution layer. In addition, Leaky ReLU [19] activation function is applied after each convolutional layer. **4**) Totally, we train 100k iterations on CASIA-B and 250k iterations on OU-MVLP. Morever, our model is optimized

Table 2. Averaged rank-1 accuracies (%) on OU-MVLP, excluding identical-view cases.

Probe	Gallery All 14 views			
	GaitSet [4]	GaitPart [6]	GLN [12]	**CSTL** (Ours)
0°	79.5	82.6	83.8	**87.1**
15°	87.9	88.9	90.0	**91.0**
30°	89.9	90.8	91.0	**91.5**
45°	90.2	91.0	91.2	**91.8**
60°	88.1	89.7	90.3	**90.6**
75°	88.7	89.9	90.0	**90.8**
90°	87.8	89.5	89.4	**90.6**
180°	81.7	85.2	85.3	**89.4**
195°	86.7	88.1	89.1	**90.2**
210°	89.0	90.0	90.5	**90.5**
225°	89.3	90.1	90.6	**90.7**
240°	87.2	89.0	89.6	**89.8**
255°	87.8	89.1	89.3	**90.0**
270°	86.2	88.2	88.5	**89.4**
Mean	87.1	88.7	89.2	**90.2**

by Adam, and the learning rate is started to set as 1e-4 and reduced to 1e-5 at 150k iterations on OU-MVLP. We use Pytorch [21] and an NVIDIA GeForce GTX 1080Ti GPU to perform our experiments.

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

CASIA-B. Tab.1 shows the comparison results between the proposed CSTL and current state-of-the-art methods in averaged rank-1 accuracies on CASIA-B dataset. Three walking conditions (NM, BG, CL) and 11 different camera views ($0^\circ - 180^\circ$) are considered into performance evaluation. Several conclusions are summarized as: 1) CSTL outperforms other methods obviously in mean accuracy comparisons under all cases, which demonstrates the robustness and advantage. 2) It's natural that performance will drop with the increase of difficulty of testing conditions. But the drop of CSTL is significantly less than other methods. Take GLN [12] as an example, the mean accuracy degradation is almost 20% (from 96.9% to 77.5%) when walking condition changes from NM to CL. Corresponding to that, the performance degradation of CSTL is 11% (from 98.0% to 87.0%). The reason is that CSTL captures the most discriminant gait features which brings the robustness to various circumstances. 3) CSTL also shows robustness to resolution of gait sequences. Comparing the performances on condition BG in two resolutions, 128×88 and 64×44, the accuracy gap for CSTL is 1.8% (from 95.4% to 93.6%), while for GaitSet the accuracy gap is 4.3% (from 91.5% to 87.2%). The improvement is still attributed to the robust feature learning of CSTL. The robustness to resolution gives CSTL another advantage that it can achieve better performance with smaller resolution in almost all cases. Based on that, we use resolution setting of 64×44 in the rest of this paper since it achieves better tradeoff between performance and computation cost.

OU-MVLP. Tab.2 shows the comparison results between the proposed CSTL and current state-of-the-art methods in averaged rank-1 accuracies on OU-MLVP. Our CSTL outperforms the existing methods under all camera views in OU-MVLP, which proves the generalization capacity of our method in a large scale dataset. It is worth noting that, CSTL is the first network which achieves average rank-1 accuracy over 90% on OU-MVLP dataset.

4.4. Ablation Study

In order to study the exact effectiveness of our method, ablation experiments are conducted to study the main components of our network. It should be noted that, our baseline does not contain any of the proposed modules in this paper.

Impact of Spatio-Temporal Modeling. The individual effects of spatial and temporal modeling are presented in Tab.3. The baseline refers to the 4-layer CNN with a feature division, while using a BA+ loss for supervision. Several notable observations can be summarized as: 1) Compared to spatial modeling network, *i.e*. GaitSet [4], our baseline achieves similar mean performance under three conditions (84.2% and 85.4%). However, with the utilization of MSTE and SSFL, our method achieves significant mean accuracy improvement over GaitSet [4] (+6.9%), which proves the superiority of our salient spatial learning capacity. 2) Compared to temporal modeling network, *i.e*., GaitPart [6], we obtain obvious improvement (from 88.0% to 90.1%) with MSTE and ATA used, which verifies the adaptive temporal representation ability in our network. 3) Applying both spatial and temporal modeling achieves the best results, which proves the complementary properties of SSFL and ATA in our method.

Impact of Multi-Scale Features. We investigate the effects of the temporal features in MSTE module and the results are given in Tab.4. It can be noticed that: 1) Comparing the first three experiments, we find that all the three level features provide positive effects on improving recognition accuracies. Thus, joint learning of the three level features achieves the best performance. 2) The inter-frame relation modelings, *i.e*., short-term and long-term, improve recognition performance based on frame-level feature learning,

Table 3. Study of the effectiveness of modules in CSTL on CASIA-B in terms of averaged rank-1 accuracy. For the sake of simplicity, we use MSTE to denote multi-scale temporal extraction.

Model	Rank-1 Accuracy			
	NM	BG	CL	Mean
GaitSet [4]	95.0	87.2	70.4	84.2
GaitPart [6]	96.2	91.5	78.7	88.0
Ours				
Baseline	95.3	88.7	72.1	85.4
Baseline + MSTE	96.6	91.1	81.0	89.6
Baseline + MSTE + ATA	97.8	93.4	79.1	90.1
Baseline + MSTE + SSFL	97.1	92.7	83.7	91.1
CSTL	**97.8**	**93.6**	**84.2**	**91.9**

Table 4. Study of the effectiveness of multi-scale temporal features on CASIA-B in terms of averaged rank-1 accuracy.

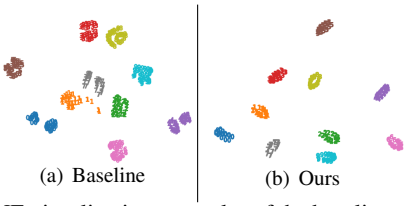
Multi-scale Features			Rank-1 Accuracy			
Frame-level	Short-term	Long-term	NM	BG	CL	Mean
✓			96.9	91.5	77.2	88.5
	✓		97.2	92.1	81.2	90.2
		✓	95.9	90.8	75.9	87.5
✓	✓		97.0	91.9	80.0	89.6
✓		✓	97.4	92.3	79.4	89.7
	✓	✓	97.4	93.2	82.0	90.9
✓	✓	✓	**97.8**	**93.6**	**84.2**	**91.9**

which proves the effectiveness on short-term and long-term temporal information. 3) Short-term and long-term features provide improvements for each other, which explains that the two type of features focus on temporal clues in complementary levels.

Table 5. Study of the effectiveness of temporal aggregation strategies on CASIA-B in terms of averaged rank-1 accuracy.

Methods	Rank-1 Accuracy			
	NM	BG	CL	Mean
Max Pooling	97.3	92.9	83.2	91.1
Average Pooling	96.8	92.3	82.7	90.6
ATA	**97.8**	**93.6**	**84.2**	**91.9**

Comparison of Sequence Aggregating Strategies. In order to investigate the effects of sequence aggregating strategy, we compare ATA with max pooling and average pooling. The results are given in Tab.5. The experimental results demonstrate the superiority of ATA. We notice that max pooling outperforms average pooling, which illustrates that extracting discriminative clues has advantages than averaging global information for fine-grained recognition task. Our ATA block outperforms max pooling and average pooling, which proves the adaptive aggregation ability of ATA.

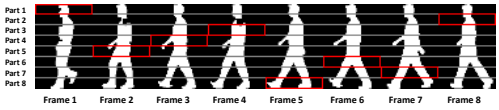


(a) Baseline (b) Ours

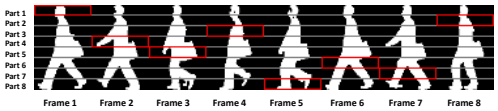
Figure 6. tSNE visualization examples of the baseline and our proposed model on CASIA-B test dataset. Different numbers with different colors indicate different identities. Best viewed with zooming in.

4.5. Visualization

We choose ten identities from CASIA-B test dataset to visualize feature distributions by t-SNE [26]. Comparing the feature distributions of baseline and our method, we notice that, in Fig.6(a), the feature distributions of different subjects are closer to each other thus identities are harder to distinguish. Differently, in Fig.6(b), the feature distributions of different subjects are more scattered to each other thus identities are more distinguishable, which proves the feature representation ability of our method.



(a) A sequence from subject '39' under NM condition with camera viewpoint of 90 degrees.



(b) A sequence from subject '106' under BG condition with camera viewpoint of 90 degrees.

Figure 7. Illustration of spatial salient feature learning. The red boxes indicate selected parts.

In order to better understand the positive effects of SSFL, we give some spatial selection examples in Fig. 7, where we set the number of selected parts as 8 in SSFL for better visualization. We can notice that: SSFL tends to select parts without body overlaps and clothing occlusions, which own complete appearance features. As shown in Fig. 7(a), SSFL selects part 8 in frame 5, which keeps the contour information of feet in a large extent compared to other frames. In Fig.7(b), under bag-carrying condition, SSFL selects part 4 in frame 2 while in other frames arms are occluded by the carrying bag. More examples are given in *supplementary material*.

In this way, we can obtain high quality spatial features, which both remedies the negative influences caused by temporal operations and enhances the robustness of our network under clothing-changing and multi-view scenarios.

5. Conclusion

In this paper, we propose a context-sensitive temporal feature learning (CSTL) network for gait recognition. CSTL extracts temporal features with multiple scales and captures salient spatial clues for achieving strong spatio-temporal modeling ability. Specifically, diverse temporal features in three scales are introduced in CSTL, and temporal relations are considered based on these temporal information for adaptive temporal aggregation. Besides, discriminative spatial parts are selected across the sequence for supplying corrupted spatial features. Extensive experiments on public datasets verify the superiority of our method.

Acknowledgements

This research is supported by the NSFC (grants No. 61773176 and No. 61733007).

References

[1] Michal Balazia and Konstantinos N Plataniotis. Human gait recognition from motion capture data in signature poses. *IET Biometrics*, 6(2):129–137, 2017. 1

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[4] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. *AAAI*, 33:8126–8133, 2019. 3, 6, 7

[5] Guangyi Chen, Yongming Rao, Jiwen Lu, and Jie Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? *European Conference on Computer Vision*, pages 660–676, 2020. 3

[6] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. *CVPR*, pages 14225–14233, 2020. 1, 3, 4, 6, 7

[7] Shreyank N Gowda, Marcus Rohrbach, and Laura Sevilla-Lara. Smart frame selection for action recognition. *arXiv preprint arXiv:2012.10671*, 2020. 5

[8] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. *European Conference on Computer Vision*, pages 228–243, 2020. 3

[9] Jinguang Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005. 3

[10] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Transactions on Information Forensics and Security*, 14(1):102–113, 2018. 3

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[12] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. *European Conference on Computer Vision*, pages 382–398, 2020. 3, 6, 7

[13] Maodi Hu, Yunhong Wang, Zhaoxiang Zhang, James J Little, and Di Huang. View-invariant discriminative projection for multi-view gait-based human identification. *IEEE Transactions on Information Forensics and Security*, 8(12):2034–2045, 2013. 3

[14] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019. 5

[15] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. *CVPR*, pages 13309–13319, 2020. 3

[16] Rijun Liao, Chunshui Cao, Edel B Garcia, Shiqi Yu, and Yongzhen Huang. Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. *Chinese conference on biometric recognition*, pages 474–483, 2017. 2

[17] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069, 2020. 2

[18] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. *ACMMM*, pages 3054–3062, 2020. 1, 3, 6

[19] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. 30(1):3, 2013. 6

[20] Ioana Macoveciuc, Carolyn J Rando, and Hervé Borrion. Forensic gait analysis and recognition: standards of evidence admissibility. *Journal of forensic sciences*, 64(5):1294–1303, 2019. 1

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS Workshops*, 2017. 7

[22] G Premalatha and Premanand V Chandramani. Improved gait recognition through gait energy image partitioning. *Computational Intelligence*, 36(3):1261–1274, 2020. 1

[23] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 2

[24] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1):4, 2018. 2, 6

[25] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. *arXiv preprint arXiv:2101.11228*, 2021. 2

[26] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[27] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. *ICIP*, pages 4165–4169, 2016. 3

[28] Haoqian Wu, Jian Tian, Yongjian Fu, Bin Li, and Xi Li. Condition-aware comparison scheme for gait recognition. *IEEE Transactions on Image Processing*, 2020. 1, 3

[29] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):209–226, 2016. 3, 6

[30] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. *ICPR*, 4:441–444, 2006. 2, 6

[31] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *TIP*, 29:1001–1015, 2019. 1, 3

[32] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. *CVPR*, pages 4710–4719, 2019. 1, 3