Self-Supervised Gait Encoding with Locality-Aware Attention for Person Re-Identification

Haocong Rao[1,3,5*] , **Siqi Wang**[2*] , **Xiping Hu**[1,4,5] , **Mingkui Tan**[3] , **Huang Da**[2] , **Jun Cheng**[1,5] and **Bin Hu**[4]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]National University of Defense Technology
[3]South China University of Technology
[4]Lanzhou University
[5]The Chinese University of Hong Kong, Hong Kong
haocongrao@gmail.com, wangsiqi10c@nudt.edu.cn, {xp.hu, jun.cheng}@siat.ac.cn
mingkuitan@scut.edu.cn, huangda1109@163.com, bh@lzu.edu.cn

Abstract

Gait-based person re-identification (Re-ID) is valuable for safety-critical applications, and using only *3D skeleton data* to extract discriminative gait features for person Re-ID is an emerging open topic. Existing methods either adopt hand-crafted features or learn gait features by traditional supervised learning paradigms. Unlike previous methods, we for the first time propose a generic gait encoding approach that can utilize *unlabeled* skeleton data to learn gait representations in a *self-supervised* manner. Specifically, we first propose to introduce self-supervision by learning to reconstruct input skeleton sequences in reverse order, which facilitates learning richer high-level semantics and better gait representations. Second, inspired by the fact that motion's continuity endows temporally adjacent skeletons with higher correlations ("*locality*"), we propose a locality-aware attention mechanism that encourages learning larger attention weights for temporally adjacent skeletons when reconstructing current skeleton, so as to learn locality when encoding gait. Finally, we propose *Attention-based Gait Encodings* (AGEs), which are built using context vectors learned by locality-aware attention, as final gait representations. AGEs are directly utilized to realize effective person Re-ID. Our approach typically improves existing skeleton-based methods by 10-20% *Rank-1* accuracy, and it achieves comparable or even superior performance to multi-modal methods with extra RGB or depth information[†].

1 Introduction

Fast, effective and reliable person re-identification (Re-ID) is vital for various applications such as security authentication,

---

[*]Authors contribute equally.
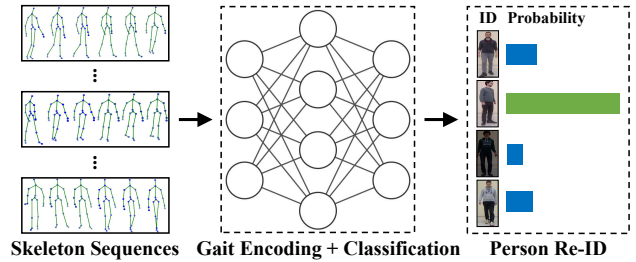[†]Codes are available at https://github.com/Kali-Hac/SGE-LA.

Figure 1: Gait-based person Re-ID using 3D skeleton data.

human tracking, and role-based activity understanding [Vezzani *et al.*, 2013]. For person Re-ID, *gait* is one of the most useful human body characteristics, and it has drawn increasing attention from researchers since it can be extracted by unobtrusive methods without co-operative subjects [Connor and Ross, 2018]. Existing physiological and psychological studies [Murray *et al.*, 1964; Cutting and Kozlowski, 1977] suggest that different human individuals have different gait characteristics, which contain numerous unique and relatively stable patterns (*e.g.*, stride length and angles of body joints). In the field of gait analysis, gait can be described by two types of methods: **(a)** *Appearance-based methods* [Zhang *et al.*, 2019], which exploit human silhouettes from aligned image sequences to depict gait. However, this type of methods are vulnerable to body shape changes and appearance changes. **(b)** *Model-based methods* [Liao *et al.*, 2020], which model gait by body structure and motion of human body joints. Unlike appearance-based methods, model-based methods are scale and view invariant [Nambiar *et al.*, 2019], thus leading to better robustness in practice. Among different models, 3D skeleton model, which describes humans by 3D coordinates of key body joints, is a highly efficient representation of human body structure and motion [Han *et al.*, 2017]. Compared with RGB or depth data, 3D skeleton enjoys many merits like better robustness and much smaller data size, and they can be easily collected by modern devices like Kinect. Therefore, exploiting 3D skeleton data to perform gait analysis for

downstream tasks like person Re-ID (illustrated in Fig. 1) has gained surging popularity [Liao *et al.*, 2020]. Nevertheless, the way to extract or learn discriminative gait features from 3D skeleton sequence data remains open to be explored.

To this end, most existing works like [Barbosa *et al.*, 2012; Andersson and Araujo, 2015] rely on hand-crafted skeleton descriptors. However, it is usually complicated and tedious to design such hand-crafted descriptors, *e.g.*, [Andersson and Araujo, 2015] defines 80 skeleton descriptors from the views of anthropometric and gait attributes for person Re-ID. Besides, they heavily rely on domain knowledge such as human anatomy [Yoo *et al.*, 2002], thus lacking the ability to mine latent gait features beyond human cognition. To alleviate this problem, few recent works like [Haque *et al.*, 2016] resort to deep neural networks to learn gait representations automatically. However, they all follow the classic *supervised* learning paradigm and require labeled data, so they cannot perform gait representation learning with unlabeled skeleton data. In this paper, we for the first time propose a ***self-supervised*** approach with ***locality-aware*** attention mechanism, which only requires unlabeled 3D skeleton data for gait encoding. First, by introducing certain self-supervision as a high-level learning goal, we not only enable our model to learn gait representations from unlabeled skeleton data, but also encourage the model to capture richer high-level semantics (*e.g.*, sequence order, body part motion) and more discriminative gait features. Specifically, we propose a self-supervised learning objective that aims to reconstruct the input skeleton sequence in reverse order, which is implemented by an encoder-decoder architecture. Second, since the continuity of motion leads to very small pose/skeleton changes in a small time interval [Aggarwal and Cai, 1999], this endows skeleton sequences with the property of locality: For each skeleton frame in a skeleton sequence, its temporally adjacent skeleton frames within a local context enjoy higher correlations to itself. Thus, to enable better skeleton reconstruction and gait representation learning, we propose the locality-aware attention mechanism to incorporate such locality into the gait encoding process. Last, we leverage the context vectors learned by the proposed locality-aware attention mechanism to construct Attention-based Gait Encodings (AGEs) as the final gait representations. We demonstrate that AGEs, which are learned without skeleton sequence labels, can be directly applied to person Re-ID and achieve highly competitive performance.

In summary, we make the following contributions:

- We propose a self-supervised approach based on reverse sequential skeleton reconstruction and an encoder-decoder model, which enable us to learn discriminative gait representations without skeleton sequence labels.

- We propose a locality-aware attention mechanism to exploit the locality nature in skeleton sequences for better skeleton reconstruction and gait encoding.

- We propose AGEs as novel gait representations, which are shown to be highly effective for person Re-ID.

Experiments demonstrate the effectiveness of our approach in gait encoding/person Re-ID. It is also shown that our model is readily transferable among datasets, which validates its abilities to learn transferable high-level features of skeletons.

2 Related Work

Person Re-ID using Skeleton Data. For person Re-ID, most existing works extract hand-crafted features to depict certain geometric, morphological or anthropometric attributes of 3D skeleton data. [Barbosa *et al.*, 2012] computes 7 Euclidean distances between the floor plane and joint or joint pairs to construct a distance matrix, which is learned by a quasi-exhaustive strategy to perform person Re-ID. [Munaro *et al.*, 2014b] further extends them to 13 skeleton descriptors (D^{13}) and leverages support vector machine (SVM) and k-nearest neighbor (KNN) for classification. In [Pala *et al.*, 2019], 16 Euclidean distances between body joints (D^{16}) are fed to Adaboost to realize Re-ID. Since skeleton based features alone are hard to achieve satisfactory Re-ID performance, features from other modalities (*e.g.*, 3D point clouds [Munaro *et al.*, 2014a], 3D face descriptor [Pala *et al.*, 2019]) are also used to enhance the performance. Meanwhile, few recent works exploit supervised deep learning models to learn gait representations from skeleton data: [Haque *et al.*, 2016] utilizes long short-term memory (LSTM) [Hochreiter and Schmidhuber, 1997] to model temporal dynamics of body joints to perform person Re-ID; [Liao *et al.*, 2020] proposes PoseGait, which feeds 81 hand-crafted pose features of 3D skeleton data into convolutional neural networks (CNN) for human recognition.

Depth-based and Multi-modal Person Re-ID Methods. Depth-based methods typically exploit human shapes or silhouettes from depth images to extract gait features for person Re-ID. [Sivapalan *et al.*, 2011] extends the Gait Energy Image (GEI) [Chunli and Kejun, 2010] to 3D domain and proposes Gait Energy Volume (GEV) algorithm based on depth images to perform gait-based human recognition. 3D point clouds based on depth data are also widely used to estimate body shape and motion trajectories. [Munaro *et al.*, 2014b] proposes point cloud matching (PCM) to compute the distances of multi-view point cloud sets, so as to discriminate different persons. [Haque *et al.*, 2016] adopts 3D LSTM to model motion dynamics of 3D point clouds for person Re-ID. As to multi-modal methods, they usually combine skeleton-based features with extra RGB or depth information (*e.g.*, depth shape features based on point clouds [Munaro *et al.*, 2014a; Hasan and Babaguchi, 2016; Wu *et al.*, 2017]) to boost Re-ID performance. In [Karianakis *et al.*, 2018], CNN-LSTM with reinforced temporal attention (RTA) is proposed for person Re-ID based on a split-rate RGB-depth transfer approach.

3 The Proposed Approach

Suppose that an input skeleton sequence $\boldsymbol{S} = (S_1, \cdots, S_f)$ contains f consecutive skeleton frames, where $S_i \in \mathbb{R}^{J \times 3}$ contains 3D coordinates of J body joints. The training set $\Phi = \{\boldsymbol{S}^{(i)}\}_{i=1}^N$ contains N skeleton sequences collected from different persons. Each skeleton sequence $\boldsymbol{S}^{(i)}$ corresponds to a label y_i, where $y_i \in \{1, \cdots, C\}$ and C is the number of persons. Our goal is to learn discriminative gait features $\boldsymbol{v}$ from $\boldsymbol{S}$ without using any label. Then, the effectiveness of learned features $\boldsymbol{v}$ is validated by using them to perform person Re-ID: Learned features and labels are used to train a simple recognition network (note that the learned
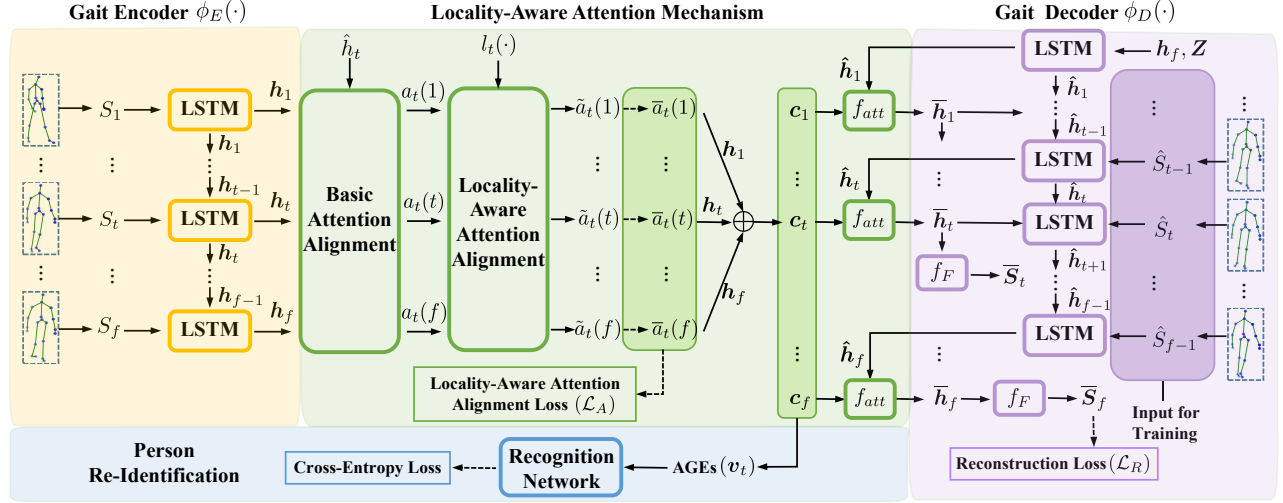
Figure 2: Flow diagram of our model: (1) Gait Encoder (yellow) encodes each skeleton frame S_t into an encoded gait state $\boldsymbol{h}_t$. (2) Locality-aware attention mechanism (green) first computes the basic attention alignment score $a_t(\cdot)$, so as to measure the content-based correlation between each encoded gait state and the decoded gait state $\hat{\boldsymbol{h}}_t$ from Gait Decoder (purple). Then, the locality mask $l_t(\cdot)$ provides an objective $\tilde{a}_t(\cdot) = a_t(\cdot) \, l_t(\cdot)$, which guides our model to learn locality-aware alignment scores $\overline{a}_t(\cdot)$ by the locality-aware attention alignment loss $\mathcal{L}_A$. Next, $\boldsymbol{h}_1 \cdots \boldsymbol{h}_f$ are weighted by $\overline{a}_t(\cdot)$ to compute the context vector $\boldsymbol{c}_t$. $\boldsymbol{c}_t$ and $\hat{\boldsymbol{h}}_t$ are fed into the concatenation layer $f_{att}(\cdot)$ to produce an attentional state vector $\overline{\boldsymbol{h}}_t$. Finally, $\overline{\boldsymbol{h}}_t$ is fed into the full connected layer $f_F(\cdot)$ to predict t^{th} skeleton $\overline{\boldsymbol{S}}_t$ and Gait Decoder for later decoding. (3) $\boldsymbol{c}_t$ is used to build Attention-based Gait Encodings (AGEs) $\boldsymbol{v}_t$, which are fed into a recognition network for person Re-ID (blue).

features $\boldsymbol{v}$ are fixed and NOT tuned by training at this stage). The overview of the proposed approach is given in Fig. 2, and we present the details of each technical component below.

3.1 Self-Supervised Skeleton Reconstruction

To learn gait representations without labels, we propose to introduce self-supervision by learning to reconstruct input skeleton sequences in *reverse* order, i.e., by taking $\boldsymbol{S}$ as inputs, we expect our model to output the sequence $\hat{\boldsymbol{S}} = (\hat{S}_1, \cdots, \hat{S}_f) = (S_f, \cdots, S_1)$, which gives $\hat{S}_t = S_{f-t+1}$. Compared with naïve reconstruction that learns to reconstruct exact inputs ($\boldsymbol{S} \longmapsto \boldsymbol{S}$), the proposed learning objective ($\boldsymbol{S} \longmapsto \hat{\boldsymbol{S}}$) is endowed with high-level information (skeleton order in a sequence) that is meaningful to human perception, which requires the model to capture richer high-level semantics (e.g., body parts, motion patterns) to achieve this learning objective. In this way, our model is expected to learn better gait representations than frequently-used plain reconstruction. Formally, given an input skeleton sequence, we use the encoder to encode each skeleton frame S_t ($t \in \{1, \cdots f\}$) and the previous step's latent state $\boldsymbol{h}_{t-1}$ (if existed), which provides context information, into current latent state $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \begin{cases} \phi_E(S_1) & \text{if} \quad t = 1 \\ \phi_E(\boldsymbol{h}_{t-1}, S_t) & \text{if} \quad t > 1 \end{cases} \quad (1)$$

where $\boldsymbol{h}_t \in \mathbb{R}^k$, $\phi_E(\cdot)$ denotes our Gait Encoder (GE). GE is built with an LSTM, which aims to capture long-term temporal dynamics of skeleton sequences. $\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f$ are *encoded gait states* that contain preliminary gait encoding information. In the **training** phase, encoded gait states are decoded by a Gait Decoder (GD) to reconstruct the target sequence $\hat{\boldsymbol{S}}$,

and the decoding process is performed below (see Fig. 2):

$$(\hat{\boldsymbol{h}}_t, \overline{S}_t) = \begin{cases} \phi_D(\boldsymbol{h}_f, \boldsymbol{Z}) & \text{if} \quad t = 1 \\ \phi_D\left(\hat{\boldsymbol{h}}_{t-1}, \hat{S}_{t-1}, \overline{\boldsymbol{h}}_{t-1}\right) & \text{if} \quad t > 1 \end{cases} \quad (2)$$

where $\phi_D(\cdot)$ denotes the GD. GD consists of an LSTM and a fully connected (FC) layer that outputs predicted skeletons. $\hat{\boldsymbol{h}}_t \in \mathbb{R}^k$ refers to the tth *decoded gait state*, i.e., the latent state output by GD's LSTM to predict t^{th} skeleton $\overline{S}_t$. When the decoding is initialized ($t = 1$), we feed an all-0 skeleton placeholder $\boldsymbol{Z} \in \mathbb{R}^J$ and the final encoded gait state $\boldsymbol{h}_f$ into GD to decode the first skeleton. Afterwards, to predict t^{th} skeleton $\overline{S}_t$, $\phi_D(\cdot)$ takes three inputs from the $(t-1)^{th}$ decoding step: decoded gait state $\hat{\boldsymbol{h}}_{t-1}$, the ground truth skeleton $\hat{S}_{t-1}$ (S_{f-t+2}) that enables better convergence, and the *attentional state vector* $\overline{\boldsymbol{h}}_{t-1}$ that fuses encoding and decoding information based on the proposed attention mechanism, which will be elaborated in Sec. 3.2. In this way, we define the objective function $\mathcal{L}_R$ for skeleton reconstruction, which minimizes the mean square error (MSE) between a target skeleton sequence $\hat{\boldsymbol{S}}$ and a predicted skeleton sequence $\overline{\boldsymbol{S}}$:

$$\mathcal{L}_R = \sum_{i=1}^{f} \sum_{j=1}^{J} (\overline{\boldsymbol{S}}_{ij} - \hat{\boldsymbol{S}}_{ij})^2 \quad (3)$$

where $\overline{\boldsymbol{S}}_{ij}, \hat{\boldsymbol{S}}_{ij}$ represent the jth joint position of the ith predicted or target skeleton. In the **testing** phase, to test the reconstruction ability of our model, we use the predicted skeleton $\overline{S}_{t-1}$ rather than target skeleton $\hat{S}_{t-1}$ as the input to ϕ_D in the $t > 1$ case, i.e., $(\hat{\boldsymbol{h}}_t, \overline{S}_t) = \phi_D\left(\hat{\boldsymbol{h}}_{t-1}, \overline{S}_{t-1}, \overline{\boldsymbol{h}}_{t-1}\right)$.

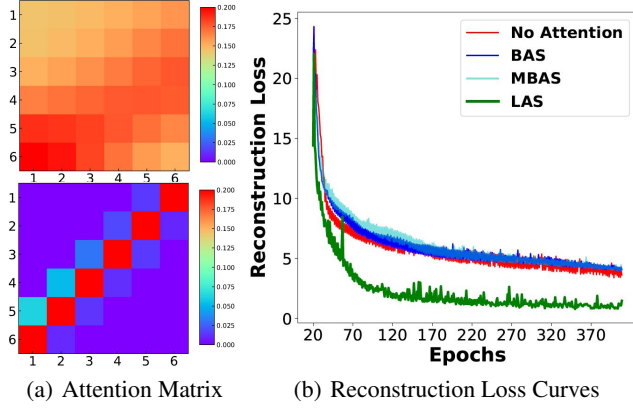(a) Attention Matrix (b) Reconstruction Loss Curves

Figure 3: (a) Visualization of the BAS (top) and LAS (bottom) attention matrices that represent average attention alignment scores. Note that the abscissa and ordinate denote indices of input skeletons and predicted skeletons respectively. (b) Reconstruction loss curves when using no attention, BAS, MBAS or LAS for reconstruction.

To facilitate training, our implementation actually optimizes Eq. 3 on each individual dimension of the skeleton's 3D coordinates: $S^{[d]} \longmapsto \hat{S}^{[d]}$, where $d \in \{X, Y, Z\}$ corresponds to a certain dimension of 3D space, and $S^{[d]}, \hat{S}^{[d]} \in \mathbb{R}^{f \times J}$.

3.2 Locality-Aware Attention Mechanism

As learning gait features requires capturing motion patterns from skeleton sequences, it is natural to consider a straightforward property of motion—Continuity. Motion's continuity ensures that those skeletons in a small temporal interval will NOT undergo drastic changes, thus resulting in higher inter-skeleton correlations in this local temporal interval, which is referred as "*locality*". Due to such locality, when reconstructing one skeleton in a sequence, we expect our model to pay more attention to its neighboring skeletons in the local context. To this end, we propose a novel locality-aware attention mechanism, with its modules presented below:

Basic Attention Alignment. We first introduce Basic Attention (BA) alignment [Luong *et al.*, 2015] to measure the content (*i.e.*, latent state) based correlations between the input sequence and the predicted sequence. As shown in Fig. 2, at t^{th} decoding step, we compute the *BA Alignment Scores (BAS)* between $\hat{h}_t$ and encoded gait state h_j ($j \in \{1, \dots, f\}$):

$$a_t(j) = \text{align}\left(\hat{h}_t, h_j\right) = \frac{\exp\left(\hat{h}_t^\top h_j\right)}{\sum_{i=1}^{f} \exp\left(\hat{h}_t^\top h_i\right)} \quad (4)$$

BAS aims to focus on those more correlative skeletons in the encoding stage, and provide preliminary attention weights for skeleton decoding. However, BA alignment only considers the content based correlations and does not explicitly take locality into consideration, which motivates us to propose locality mask and locality-aware attention alignment below.

Locality Mask. Our motivation is to incorporate locality into the gait encoding process for better skeleton reconstruction. As the goal is to decode the tth skeleton $\overline{S}_t$ as $\hat{S}_t$, we

assume those skeletons in the local temporal context of S_{p_t} to be highly correlated to $\overline{S}_t$, where $p_t = f - t + 1$ (note that we use reverse reconstruction). To describe the local context centered at S_{p_t}, we define an attentional window $[p_t - D, p_t + D]$, where D is a selected integer to control the attentional range. Since the locality will favor temporal positions near p_t (closer positions are more correlative), a direct solution is to place a *Gaussian* distribution centered around p_t as a locality mask:

$$l_t(j) = \exp\left(-\frac{(j - p_t)^2}{2\sigma^2}\right) \quad (5)$$

where we empirically set $\sigma = \frac{D}{2}$, j is a position within the window centered at p_t. We can weight BAS by this locality mask to compute *Masked BA Alignment Scores (MBAS)* below, which directly forces alignment scores to obtain locality:

$$\tilde{a}_t(j) = l_t(j) \cdot a_t(j) \quad (6)$$

Locality-Aware Attention Alignment. Despite that the locality mask is straightforward to yield locality, it is a very coarse solution that brutally changes the alignment scores. Therefore, instead of using MBAS ($\tilde{a}_t(j)$) directly, we propose the *Locality-aware Attention (LA) alignment*. Specifically, an LA alignment loss term $\mathcal{L}_A$ is used to encourage LA alignment to learn similar locality of the $\tilde{a}_t(j)$:

$$\mathcal{L}_A = \sum_{t=1}^{f} \sum_{j=1}^{f} (a_t(j) - \tilde{a}_t(j))^2 \quad (7)$$

By adding the loss term $\mathcal{L}_A$, we can obtain *LA Alignment Scores (LAS)*. Note that in Eq. 4, the final learned $a_t(j)$ is BAS. For clarity, we use $\overline{a}_t(j)$ to represent learned LAS here.

With the guidance of $\mathcal{L}_A$, our model learns to allocate more attention to the local context by itself, rather than using a locality mask. To utilize alignment scores to yield an attention-weighted encoded gait state at the tth step, we calculate the *context vector* c_t by a sum of weighted encoded gait states:

$$c_t = \sum_{j=1}^{f} \overline{a}_t(j) h_j \quad (8)$$

Note that context vector c_t can also be computed by BAS or MBAS. c_t provides a synthesized gait encoding that is more relevant to $\hat{h}_t$, which facilitates the reconstruction of tth skeleton. To combine both encoding and decoding information for reconstruction, we use a concatenation layer $f_{att}(\cdot)$ that combines c_t and $\hat{h}_t$ into an attentional state vector $\overline{h}_t$:

$$\overline{h}_t = f_{att}(c_t, \hat{h}_t) = \tanh(W_{att}[c_t; \hat{h}_t]) \quad (9)$$

where W_{att} represents the learnable weight matrix in the layer. Finally, we generate (reconstruct) the tth skeleton by the FC layer $f_F(\cdot)$ of the GD:

$$\overline{S}_t = f_F\left(\overline{h}_t\right) = W_F \overline{h}_t \quad (10)$$

where W_F is the weights to be learned in this FC layer.

Remarks: (a) To provide an intuitive illustration of the proposed locality-aware attention mechanism, we visualize the BAS and LAS attention matrices, which are formed by

average alignment scores computed on dimension X as an example. As shown by Fig. 3(a), LA alignment significantly improves the locality of learned alignment scores: Relatively large alignment scores are densely distributed near the clinodiagonal line of the alignment score matrix (clinodiagonal line reflects skeletons' correlations to themselves when performing reverse reconstruction), which means adjacent skeletons are assigned with larger attention than remote skeletons. By contrast, despite that BA alignment also learns locality to some extent, BA's alignment weights show a much more even distribution, and many non-adjacent skeletons are also given large alignment scores. Similar trends are also observed for dimension Y and Z. **(b)** The proposed LA alignment enables us to learn better reconstruction of skeleton sequences. We visualize the reconstruction loss during training in four cases: using no attention mechanism, BAS, MBAS and LAS. As shown by Fig. 3(b), it can be observed that training with LAS converges at a faster speed with an evidently smaller reconstruction loss, which justifies our intuition that locality will facilitate training. Interestingly, we observe that using the locality mask directly in fact does not benefit training, and it verifies that learning is a better way to obtain locality.

3.3 Attention-based Gait Encodings (AGEs)

Since our ultimate goal is to learn good gait features from skeleton data to perform person Re-ID, we need to extract certain internal skeleton embedding from the proposed model as gait representations. Unlike traditional LSTM based methods that basically rely on the last hidden state to compress the temporal dynamics of a sequence [Weston *et al.*, 2015], we recall that the dynamic context vector $\mathbf{c}_t$ from the attention mechanism integrates the key encoded gait states of input skeletons and retains crucial spatio-temporal information to reconstruct target skeletons. Hence, we utilize them instead of the last hidden state to build our final gait representations—AGEs. A skeleton-level AGE ($\boldsymbol{v}_t$) is defined as follows:

$$\boldsymbol{v}_t = [\boldsymbol{c}_t^X; \boldsymbol{c}_t^Y; \boldsymbol{c}_t^Z] \qquad (11)$$

where $\boldsymbol{c}_t^d$ denotes the context vector computed on dimension $d \in \{X, Y, Z\}$ of the tth step in decoding. To perform person Re-ID, we use $\boldsymbol{v}_t$ to train a simple recognition network $f_{RN}(\cdot)$ that consists of a hidden layer and a softmax layer. We average the prediction of each skeleton-level AGE $f_{RN}(\boldsymbol{v}_t; \theta_\tau)$ ($t \in \{1, \cdots, f\}$) in a skeleton sequence to be the final sequence-level prediction for person Re-ID, where θ_τ refers to parameters of $f_{RN}(\cdot)$. Note that during training, each skeleton in one sequence shares the same skeleton sequence label y_i. Besides, skeleton labels are only used to train the recognition network, *i.e.*, AGEs are fixed during training to demonstrate the effectiveness of learned gait features.

3.4 The Entire Approach

As a summary, the computation flow of the entire approach during skeleton reconstruction is $\boldsymbol{h} \to \hat{\boldsymbol{h}} \to \overline{a} \to \boldsymbol{c} \to \overline{\boldsymbol{h}} \to \overline{\boldsymbol{S}}$. To guide model training in the reconstruction process, we combine the skeleton reconstruction loss $\mathcal{L}_R$ in Eq. 3 and the LA alignment loss $\mathcal{L}_A$ in Eq. 7 as follows:

$$\mathcal{L} = \lambda_R \mathcal{L}_R + \lambda_A \mathcal{L}_A + \beta \left\| \Theta \right\|_2^2 \qquad (12)$$

where Θ denotes the parameters of the model, λ_R, λ_A are the weight coefficients to trade off the importance of the reconstruction loss and LA alignment loss, $\left\| \Theta \right\|_2^2$ is L_2 regularization. For the person Re-ID task, we employ the cross-entropy loss to train the recognition network with AGEs ($\boldsymbol{v}_t$).

4 Experiments

4.1 Experimental Settings

Datasets: We evaluate our method on three public Re-ID datasets that provide 3D skeleton data: BIWI [Munaro *et al.*, 2014b], IAS-Lab [Munaro *et al.*, 2014c] and Kinect Gait Biometry Dataset (KGBD) [Andersson and Araujo, 2015]. They collect skeleton data from 50, 11 and 164 different individuals respectively. We follow the evaluation setup in [Haque *et al.*, 2016], which is frequently used in the literature: For BIWI, we use the full training set and the Walking testing set that contains dynamic skeleton data; For IAS-Lab, we use the full training set and two test splits, IAS-A and IAS-B; For KGBD, since no training and testing splits are given, we randomly leave one skeleton video of each person for testing and use the rest of videos for training. The experiments are repeated for multiple times and the average performance is reported on KGBD. We discard the first and last 10 frames of each original skeleton sequence to avoid ineffective skeleton recording. Then, we spilt the training dataset into multiple skeleton sequences with the length f, and two consecutive sequences share $\frac{f}{2}$ overlapping skeletons, which aims to obtain as many skeleton sequences as possible to train our model.

Implementation Details: The number of body joints is set to the maximum number in all datasets, namely $J = 20$. The sequence length f is empirically set to 6 as it achieves the best performance in average among different sequence length settings. To learn the locality-aware attention for the whole sequence, the attentional range D of LA is set to 6. We use a 2-layer LSTM with $k = 256$ hidden units per layer for both GE and GD. We empirically set both λ_R and λ_A to 1, while a momentum 0.9 is utilized for optimization. We use a learning rate $lr = 0.0005$, and we set the weight of L_2 regularization β to 0.02. The batch size is set to 128 in all experiments.

Evaluation Metrics: Person Re-ID typically follows a "multi-shot" manner that leverages predictions of multiple frames or a sequence-level representation to produce a sequence label. In this paper, we compute both *Rank-1* accuracy and *nAUC* (area under the cumulative matching curve (CMC) normalized by the number of ranks [Gray and Tao, 2008]) to evaluate multi-shot person Re-ID performance.

4.2 Performance Comparison

In Table 1, we conduct an extensive comparison with existing skeleton based person Re-ID methods (Id = 10-15) in the literature. In the meantime, we also include classic depth-based methods (Id = 1-3) and representative multi-modal methods (Id = 4-9) as a reference. We obtain observations below:

Comparison with Skeleton-based Methods: As shown by Table 1, our approach enjoys obvious advantages over existing skeleton-based methods in terms of person Re-ID performance: First, our approach evidently outperforms those

Id	Methods	Rank-1 (%)				nAUC			
		BIWI	IAS-A	IAS-B	KGBD	BIWI	IAS-A	IAS-B	KGBD
1	Gait Energy Image [2010]	21.4	25.6	15.9	—	73.2	72.1	66.0	—
2	Gait Energy Volume [2011]	25.7	20.4	13.7	—	83.2	66.2	64.8	—
3	3D LSTM [2016]	27.0	31.0	33.8	—	83.3	77.6	78.0	—
4	PCM + Skeleton [2014a]	42.9	27.3	81.8	—	—	—	—	—
5	Size-Shape decriptors + SVM [2016]	20.5	—	—	—	—	—	—	—
6	Size-Shape decriptors + LDA [2016]	22.1	—	—	—	—	—	—	—
7	DVCov + SKL [2017]	21.4	46.6	45.9	—	—	—	—	—
8	ED + SKL [2017]	30.0	52.3	63.3	—	—	—	—	—
9	CNN-LSTM with RTA [2018]	50.0	—	—	—	—	—	—	—
10	D^{13} descriptors + SVM [2014b]	17.9	—	—	—	—	—	—	—
11	D^{13} descriptors + KNN [2014b]	39.3	33.8	40.5	46.9	64.3	63.6	71.1	90.0
12	D^{16} descriptors + Adaboost [2019]	41.8	27.4	39.2	69.9	74.1	65.5	78.2	90.6
13	Single-layer LSTM [2016]	15.8	20.0	19.1	39.8	65.8	65.9	68.4	87.2
14	Multi-layer LSTM [2019]	36.1	34.4	30.9	46.2	75.6	72.1	71.9	89.8
15	PoseGait [2020]	33.3	41.4	37.1	**90.6**	81.8	79.9	74.8	**97.8**
16	**Ours**	**59.1**	**56.1**	**58.2**	87.7	**86.5**	**81.7**	**85.3**	96.3

Table 1: Comparison with existing skeleton-based methods (10-15). Depth-based methods (1-3) and multi-modal methods (4-9) are also included as a reference. Bold numbers refer to the best performers among skeleton-based methods. "—" indicates no published result.

methods that rely on manually-designed geometric or anthropometric skeleton descriptors (Id = 10-12). For example, D^{13} (Id = 11) and recent D^{16} (Id = 12) are two most representative hand-crafted feature based methods, and our model outperforms both of them by a large margin (17.3%-40.8% *Rank-1* accuracy and 5.7%-22.2% *nAUC* on different datasets). Second, our approach is also superior to recent skeleton based methods that utilize deep neural networks (Id = 13-15). Our approach is the best performer on three out of four datasets (BIWI, IAS-A, IAS-B) with a 14.7%-43.3% *Rank-1* accuracy and 1.8%-20.7% *nAUC* gain. On KGBD dataset, our approach ranks 2nd and performs slightly inferior to the latest PoseGait (Id = 15). However, despite that CNN is used, PoseGait still requires extracting 81 hand-crafted features. Besides, labeled skeleton data are indispensable for existing deep learning based methods, while our approach can learn better gait representations by unlabeled skeletons only.

Comparison with Depth-based Methods and Multimodal Methods: Despite that our approach only takes skeleton data as inputs, our approach consistently outperforms baselines of classic depth-based methods (Id = 1-3) by at least 24% *Rank-1* and 3.2% *nAUC* gain. Considering the fact that skeleton data are of much smaller size than depth image data, our approach is both effective and efficient. As to the comparison with recent methods that exploit multi-modal inputs (Id = 4-9), the performance of our approach is still highly competitive: Although few multi-modal methods perform better on IAS-B, our skeleton based method achieves the best *Rank-1* accuracy on BIWI and IAS-A. Interestingly, we note that the multi-modal approach that uses both point cloud matching (PCM) and skeletons yields the best accuray on IAS-B, but it performs markedly worse on datasets that undergo more frequent shape and appearance changes (IAS-A and BIWI). By contrast, our approach consistently achieves stable and satisfactory performance on each dataset. Thus, with 3D skeletons as the sole input, our approach can be a promising solution to person Re-ID and other potential skeleton-related tasks.

GE	GD	Rev.	BAS	MBAS	LAS	AGEs	Rank-1	nAUC
✓							36.1	75.6
✓	✓						41.5	80.1
✓	✓	✓					46.7	81.5
✓	✓		✓			✓	45.7	84.1
✓	✓	✓	✓				53.3	84.6
✓	✓	✓	✓			✓	55.1	85.2
✓	✓			✓			52.9	85.0
✓	✓			✓		✓	53.1	83.6
✓	✓	✓		✓		✓	54.5	85.6
✓	✓	✓			✓		57.7	85.8
✓	✓				✓	✓	57.2	85.7
✓	✓	✓			✓	✓	**59.1**	**86.5**

Table 2: Ablation study of our model. "✓" indicates that the corresponding model component is used: GE, GD, reverse skeleton reconstruction (Rev.), different types of attention alignment scores (BAS, MBAS, LAS). "AGEs" indicates exploiting AGEs (v_t) rather than encoded gait states of GE's LSTM h_t for person Re-ID.

5 Discussion

Ablation Study. We perform ablation study to verify the effectiveness of each model component. As shown in Table 2, we draw the following conclusions: **(a)** The proposed encoder-decoder architecture (GE-GD) performs remarkably better (5.4% *Rank-1* accuracy and 4.5% *nAUC* gain) than supervised learning paradigm that uses GE only, which verifies the necessity of encoder-decoder architecture and skeleton reconstruction mechanism. **(b)** Using reverse reconstruction (Rev. in Table 2) produces evident performance gain (1.4%-9.4% *Rank-1* accuracy and 0.8%-1.4% *nAUC*) when compared with those configurations without reverse reconstruction. Such results justify our claim that reverse reconstruction enables the model to learn more discriminative gait features for person Re-ID. **(c)** Introducing attention mechanism (BAS) improves the model by 6.6% *Rank-1* accuracy and 3.1% *nAUC*, while LAS further improves BAS's performance by 4.4% *Rank-1* and 1.2% *nAUC*. Besides, it is noted that directly using locality mask (MBAS) degrades the per-

Config.	BIWI		IAS-A		IAS-B	
	Self	Transfer	Self	Transfer	Self	Transfer
BAS	**55.1**	53.5	**54.7**	54.2	56.3	**57.1**
LAS	**59.1**	58.4	56.1	**56.3**	**58.2**	57.4

Table 3: *Rank-1* accuracy comparison between the original model ("Self") and the transferred model ("Transfer"). Results of different datasets and alignment score types (BAS or LAS) are reported.

formance. This substantiates our claim that learning locality enables better gait representation learning. **(d)** AGEs provide more effective gait representations: Using AGEs can consistently improve the Re-ID performance by up to 1.8% *Rank-1* accuracy and 0.7% *nAUC*, regardless of the type of used alignment scores. Other datasets report similar results.

Gait Encoding Model Transfer. We discover that the gait encoding model learned on one dataset can be readily transferred to other datasets. For example, we use the model pre-trained on KGBD to directly encode skeletons from BIWI and IAS-Lab, and compare the Re-ID performance with the model trained on BIWI and IAS-Lab themselves (denoted by "Self"): As shown by Table 3, two cases achieve fairly comparable performance, while the transferred model even outperforms the original model on IAS-A (LAS) and IAS-B (BAS). Such transferability demonstrates that our approach can learn transferable high-level semantics of 3D skeleton data, which enables the pre-trained model to capture discriminative gait features from unseen skeletons of a new dataset.

6 Conclusion

In this paper, we propose a generic self-supervised approach to learn effective gait representations for person Re-ID. We introduce self-supervision by learning reverse skeleton reconstruction, which enables our model to learn high-level semantics and discriminative gait features from unlabeled skeleton data. To facilitate skeleton reconstruction and gait representation learning, a novel locality-aware attention mechanism is proposed to incorporate the locality into gait encoding process. We construct AGEs as final gait representations to perform person Re-ID. Our approach evidently outperforms existing skeleton-based Re-ID methods, and its performance is comparable or superior to depth-based/multi-modal methods.

Acknowledgements

References

[Aggarwal and Cai, 1999] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[Andersson and Araujo, 2015] Virginia Ortiz Andersson and Ricardo Matsumura Araujo. Person identification using anthropometric and gait data from kinect sensor. In *AAAI*, 2015.

[Barbosa *et al.*, 2012] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV*, pages 433–442. Springer, 2012.

[Chunli and Kejun, 2010] Lin Chunli and Wang Kejun. A behavior classification based on enhanced gait energy image. In *2nd International Conference on Networking and Digital Society*, volume 2, pages 589–592. IEEE, 2010.

[Connor and Ross, 2018] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer vision and image understanding*, 167:1–27, 2018.

[Cutting and Kozlowski, 1977] James E Cutting and Lynn T Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.

[Gray and Tao, 2008] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275. Springer, 2008.

[Han *et al.*, 2017] Fei Han, Brian Reily, William Hoff, and Hao Zhang. Space-time representation of people based on 3d skeletal data: A review. *Computer Vision and Image Understanding*, 158:85–105, 2017.

[Haque *et al.*, 2016] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, pages 1229–1238, 2016.

[Hasan and Babaguchi, 2016] Mohamed Hasan and Noborou Babaguchi. Long-term people reidentification using anthropometric signature. In *8th International Conference on Biometrics Theory, Applications and Systems*, pages 1–6. IEEE, 2016.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Karianakis *et al.*, 2018] Nikolaos Karianakis, Zicheng Liu, Yinpeng Chen, and Stefano Soatto. Reinforced temporal attention and split-rate transfer for depth-based person re-identification. In *ECCV*, pages 715–733, 2018.

[Liao *et al.*, 2020] 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.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[Munaro *et al.*, 2014a] Matteo Munaro, Alberto Basso, Andrea Fossati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *ICRA*, pages 4512–4519. IEEE, 2014.

[Munaro *et al.*, 2014b] Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti, and Luc Van Gool. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, pages 161–181. Springer, 2014.

[Munaro *et al.*, 2014c] Matteo Munaro, Stefano Ghidoni, Deniz Tartaro Dizmen, and Emanuele Menegatti. A feature-based approach to people re-identification using skeleton keypoints. In *ICRA*, pages 5644–5651. IEEE, 2014.

[Murray *et al.*, 1964] M Pat Murray, A Bernard Drought, and Ross C Kory. Walking patterns of normal men. *JBJS*, 46(2):335–360, 1964.

[Nambiar *et al.*, 2019] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys (CSUR)*, 52(2):33, 2019.

[Pala *et al.*, 2019] Pietro Pala, Lorenzo Seidenari, Stefano Berretti, and Alberto Del Bimbo. Enhanced skeleton and face 3d data for person re-identification from depth cameras. *Computers & Graphics*, 79:69–80, 2019.

[Sivapalan *et al.*, 2011] Sabesan Sivapalan, Daniel Chen, Simon Denman, Sridha Sridharan, and Clinton Fookes. Gait energy volumes and frontal gait recognition using depth images. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2011.

[Vezzani *et al.*, 2013] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys (CSUR)*, 46(2):29, 2013.

[Weston *et al.*, 2015] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *ICLR*, 2015.

[Wu *et al.*, 2017] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017.

[Yoo *et al.*, 2002] Jang-Hee Yoo, Mark S Nixon, and Chris J Harris. Extracting gait signatures based on anatomical knowledge. In *Proceedings of BMVA Symposium on Advancing Biometric Technologies*, pages 596–606. Citeseer, 2002.

[Zhang *et al.*, 2019] Yuqi Zhang, Yongzhen Huang, Liang Wang, and Shiqi Yu. A comprehensive study on gait biometrics using a joint cnn-based method. *Pattern Recognition*, 93:228–236, 2019.

[Zheng *et al.*, 2019] Wu Zheng, Lin Li, Zhaoxiang Zhang, Yan Huang, and Liang Wang. Relational network for skeleton-based action recognition. In *ICME*, pages 826–831. IEEE, 2019.