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], 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

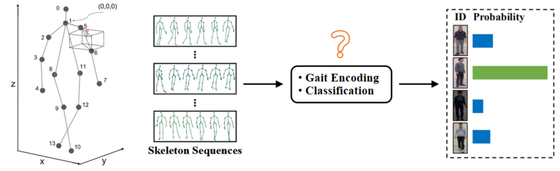
[4] Lanzhou University, [5] The Chinese University of Hong Kong, Hong Kong

Objectives

Skeleton-based person re-identification (Re-ID) aims to use 3D skeletons to re-identify the same person in a different view.

Task and Approach

- Encode geometric, anthropometirc, gait attributes
- ID classification based on the learned representation

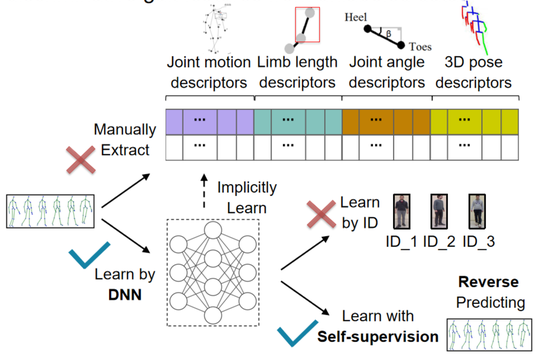


Example of 3D skeleton Objective of 3D skeleton-based Re-ID task

- We introduce a novel *self-supervised* approach based on **reverse sequential reconstruction** and **locality-aware attention** to learn effective gait representation.

Motivation

- No need of hand-crafted descriptors or labeled data
- Automatically learn high-level semantics and discriminative gait features from unlabeled data



Proposed Approach

- **Self-Supervised Reverse Skeleton Reconstruction with Locality-Aware Attention Mechanism**

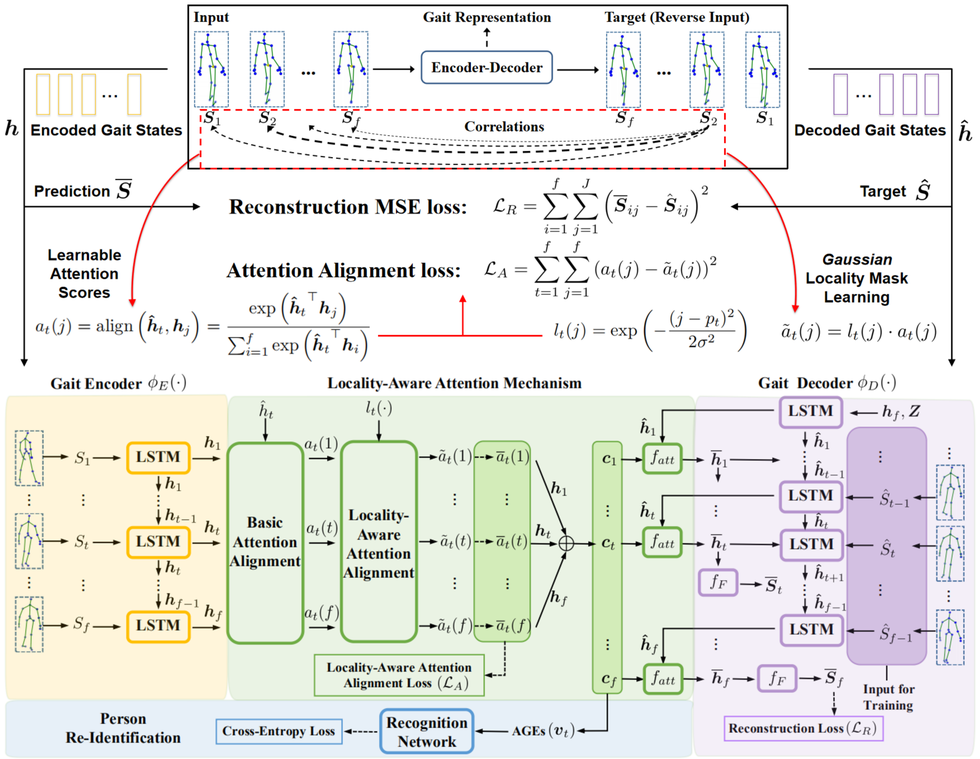


Figure 2: Flow diagram of our model: (1) Gait Encoder (yellow) encodes each skeleton frame S_t into an encoded gait state $\boldsymbol{h}_t$. (2) Locality-aware attention mechanism (green) first computes the basic attention alignment score $a_t(\cdot)$, so as to measure the content-based correlation 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).

Experiements

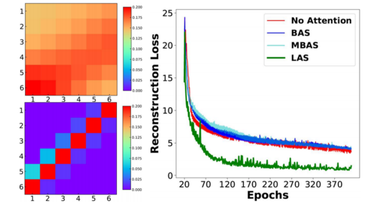
- **Comparisons results and ablation study**

	Id	Methods	Rank-1 (%) BIWI	IAS-A	IAS-B	KGBD	nAUC BIWI	IAS-A	IAS-B	KGBD
Depth-based methods	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	—
Multi-modal methods	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	—	—	—	—	—	—	—
Skeleton-based methods	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

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 ($\boldsymbol{v}_t$) rather than encoded gait states of GE's LSTM $\boldsymbol{h}_t$ for person Re-ID.

- **(a) LAS improves the locality of learned scores**
- **(b) LAS significantly reduces reconstruction loss**



(a) Attention Matrix (b) Reconstruction Loss Curves

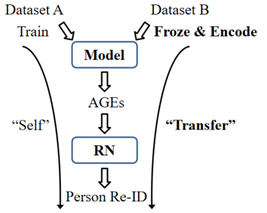


Illustration for gait encoding model "transfer"

- **Our model learns high-level semantics of 3D skeleton data**

⇨ **Pre-trained models are transferrable**

Config.	BIWI Self	BIWI Transfer	IAS-A Self	IAS-A Transfer	IAS-B Self	IAS-B 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.

Related Links

- Models and codes: https://github.com/Kali-Hac/SGE-LA
- Latest extended work: https://arxiv.org/pdf/2009.03671