SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification

Haocong Rao and **Chunyan Miao**†

School of Computer Science and Engineering, Nanyang Technological University
Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY)
haocong001@ntu.edu.sg, ascymiao@ntu.edu.sg

Abstract

Recent advances in skeleton-based person re-identification (re-ID) obtain impressive performance via either hand-crafted skeleton descriptors or skeleton representation learning with deep learning paradigms. However, they typically require skeletal pre-modeling and label information for training, which leads to limited applicability of these methods. In this paper, we focus on *unsupervised* skeleton-based person re-ID, and present a generic Simple Masked Contrastive learning (SimMC) framework to learn effective representations from *unlabeled* 3D skeletons for person re-ID. Specifically, to fully exploit skeleton features within each skeleton sequence, we first devise a *masked prototype contrastive learning (MPC)* scheme to cluster the most typical skeleton features (*skeleton prototypes*) from different subsequences randomly masked from raw sequences, and contrast the inherent similarity between skeleton features and different prototypes to learn discriminative skeleton representations without using any label. Then, considering that different subsequences within the same sequence usually enjoy strong correlations due to the nature of motion continuity, we propose the *masked intra-sequence contrastive learning (MIC)* to capture intra-sequence pattern consistency between subsequences, so as to encourage learning more effective skeleton representations for person re-ID. Extensive experiments validate that the proposed SimMC outperforms most state-of-the-art skeleton-based methods. We further show its scalability and efficiency in enhancing the performance of existing models. Our codes are available at https://github.com/Kali-Hac/SimMC.

1 Introduction

Person re-identification (re-ID) targets at retrieving and matching the same pedestrian from different views or occasions, which assumes a pivotal role in various applications such as intelligent surveillance, robotics, and security authentication [Ye *et al.*, 2021]. Recently, person re-ID via 3D skeletons has drawn growing interests from academia and industry [Pala *et al.*, 2019; Rao *et al.*, 2021b; Rao *et al.*, 2021d]. Compared with conventional image-based methods that typically rely on visual features such as human silhouettes and appearances for recognition [Liu *et al.*, 2015], skeleton-based methods leverage 3D positions of key body joints to characterize discriminative structural and motion features of human body, which could enjoy smaller data size and better robustness against scale and view variation [Han *et al.*, 2017].

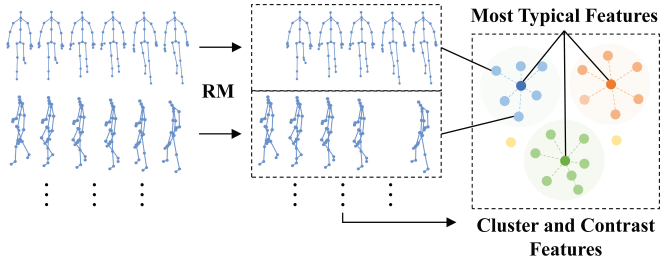


Figure 1: Our framework clusters the randomly masked (RM) skeleton sequences, and contrasts their features with the most typical ones to learn discriminative skeleton representations for person re-ID.

Despite the great progress in skeleton-based person re-ID, existing endeavors require either extracting hand-crafted features ($e.g.$, anthropometric attributes) [Pala *et al.*, 2019] or learning skeleton representations with the supervision of labels. For hand-crafted methods, they typically require extensive domain knowledge while lacking the flexibility to explore latent features beyond human cognition. To tackle this issue, numerous recent works resort to convolutional neural networks (CNN) [Liao *et al.*, 2020] and long short-term memory (LSTM) [Rao *et al.*, 2021a] to perform supervised or self-supervised skeleton representation learning. However, these methods usually require a specific pre-modeling of 3D skeletons ($e.g.$, skeleton graphs [Rao *et al.*, 2021d]), and rely on massive manually-annotated data to train or fine-tune models, which is labor-expensive and unable to learn general pedestrian representations under the unavailability of labels.

To address these challenges, this paper presents a generic Simple Masked Contrastive learning (SimMC) framework, as shown in Fig. 1, which contrasts the typical features and in-

†Corresponding author

herent relationships of *masked* skeleton sequences to learn effective skeleton representations *without using any label* for person re-ID. Specifically, to fully utilize unique features within skeleton sequences, we first devise a ***masked prototype contrastive learning (MPC)*** scheme to cluster *subsequence* representations (referred as *skeleton instances*) randomly masked from raw sequences, and contrast the inherent similarity between them and the most typical features (referred as *skeleton prototypes*) to learn discriminative skeleton representations. By pulling closer skeleton instances belonging to the same prototype and pushing apart instances of different prototypes with the instance-prototype contrastive learning, MPC enables the model to capture discriminative skeleton features and high-level semantics ($e.g.$, intra-class skeleton similarity) from *unlabeled* skeleton sequences for the person re-ID task. Then, motivated by the nature of motion continuity that typically endows different subsequences with strong correlations ($e.g.$, motion similarity), we propose the ***masked intra-sequence contrastive learning (MIC)*** to learn the intra-sequence similarity between subsequences of the same skeleton sequence, which encourages capturing the pattern consistency within sequences to learn more effective representations of skeletons for person re-ID.

The proposed SimMC framework enjoys merits in terms of architectures, performance, and scalability. Firstly, SimMC is primarily built by multi-layer perceptron (MLP) networks with small model complexity, which can directly learn effective representations from raw skeleton sequences without any prior modeling. Secondly, the proposed unsupervised framework outperforms most existing self-supervised and supervised skeleton-based methods that utilize extra label information, and can also be efficiently applied to 3D skeleton data estimated from RGB-based scenes. Lastly, our framework can serve as a generic contrastive learning paradigm to fine-tune skeleton features learned from existing models, which benefits learning better skeleton representations for the task of person re-ID. In summary, our main contributions include:

- We present a simple masked contrastive learning (SimMC) framework that exploits typical features and relationships of masked unlabeled skeleton sequences to learn discriminative representations for person re-ID.
- We devise a novel masked prototype contrastive learning (MPC) scheme to fully contrast most representative features and learn high-level semantics from subsequence representations masked from skeleton sequences.
- We propose the masked intra-sequence contrastive learning (MIC) to learn inherent similarity and pattern consistency between subsequences, so as to encourage learning more effective representations for person re-ID.
- Empirical evaluations show that SimMC significantly outperforms most state-of-the-art skeleton-based methods on four benchmark datasets, and can be exploited to fine-tune existing skeleton representations and boost their performance with up to 28.2% mAP gains.

2 Related Works

Skeleton-based Person Re-identification. Most existing methods typically extract hand-crafted anthropometric, morphological, and gait descriptors from 3D skeletons to characterize human body and motion features. Seven Euclidean distances between certain joints are utilized by [Barbosa *et al.*, 2012] to construct a distance matrix for person re-ID. Further enhancement with 13 (D_{13}) and 16 skeleton descriptors (D_{16}) are made in [Munaro *et al.*, 2014a] and [Pala *et al.*, 2019], respectively, which leverage k-nearest neighbor, support vector machine or Adaboost classifiers to perform person re-ID. Recently, deep neural networks are widely applied to supervised and self-supervised skeleton representation learning. A CNN-based paradigm, PoseGait [Liao *et al.*, 2020], is devised to encode 81 hand-crafted skeleton/pose features for human recognition. An LSTM-based skeleton encoding model with locality-aware attention (AGE) [Rao *et al.*, 2020] is proposed to learn discriminative gait features from skeleton sequences. SGELA [Rao *et al.*, 2021b] further combines multiple self-supervised pretext tasks ($e.g.$, reverse sequential reconstruction) and inter-sequence contrastive scheme to enhance skeleton pattern learning for person re-ID. The graph-based methods MG-SCR [Rao *et al.*, 2021d] and SM-SGE [Rao *et al.*, 2021a] devise multi-level skeleton graphs and auxiliary self-supervised tasks for person re-ID tasks.

Contrastive Learning. Contrastive learning is widely applied to various self-supervised and unsupervised paradigms [He *et al.*, 2020; Rao *et al.*, 2021b; Rao *et al.*, 2021c; Chen and He, 2021] to learn effective data representations by pulling together positive representation pairs and pushing apart negative ones in a certain feature space. An instance discrimination paradigm based on exemplar tasks [Wu *et al.*, 2018] is devised for image contrastive learning. The contrastive predictive coding (CPC) model with the probabilistic InfoNCE loss [Oord *et al.*, 2018] is proposed to learn general representations from various domains. Recent contrastive paradigms explore mini-batch negative sampling [Chen *et al.*, 2020] and momentum-based encoders [He *et al.*, 2020], while [Chen and He, 2021] devises a Siamese architecture for contrastive learning without using negative pairs or momentum encoders. In [Li *et al.*, 2021], contrastive learning and k-means clustering are combined for unsupervised learning of visual representations.

3 The Proposed Framework

Suppose that a 3D skeleton sequence $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f) \in \mathbb{R}^{f \times K}$, where $\boldsymbol{S}_t \in \mathbb{R}^K$ is the t^{th} skeleton with 3D coordinates of J body joints and $K = J \times 3$. Each skeleton sequence $\boldsymbol{S}_{1:f}$ belongs to an identity y, where $\mathrm{y} \in \{1, \cdots, I\}$ and I is the number of different identities. The training set $\Phi_{\mathcal{T}} = \left\{\boldsymbol{S}_{1:f}^{\mathcal{T},i}\right\}_{i=1}^{N_1}$, probe set $\Phi_{\mathcal{P}} = \left\{\boldsymbol{S}_{1:f}^{\mathcal{P},i}\right\}_{i=1}^{N_2}$, and gallery set $\Phi_{\mathcal{G}} = \left\{\boldsymbol{S}_{1:f}^{\mathcal{G},i}\right\}_{i=1}^{N_3}$ contain N_1, N_2, and N_3 skeleton sequences of different persons in different views and scenes. Our framework aims at learning an encoder (denoted as $\psi(\cdot)$) built with neural networks to encode $\Phi_{\mathcal{P}}$ and $\Phi_{\mathcal{G}}$ into effective skeleton representations $\{\boldsymbol{v}_i^{\mathcal{P}}\}_{i=1}^{N_2}$ and $\{\boldsymbol{v}_j^{\mathcal{G}}\}_{j=1}^{N_3}$ *without using*

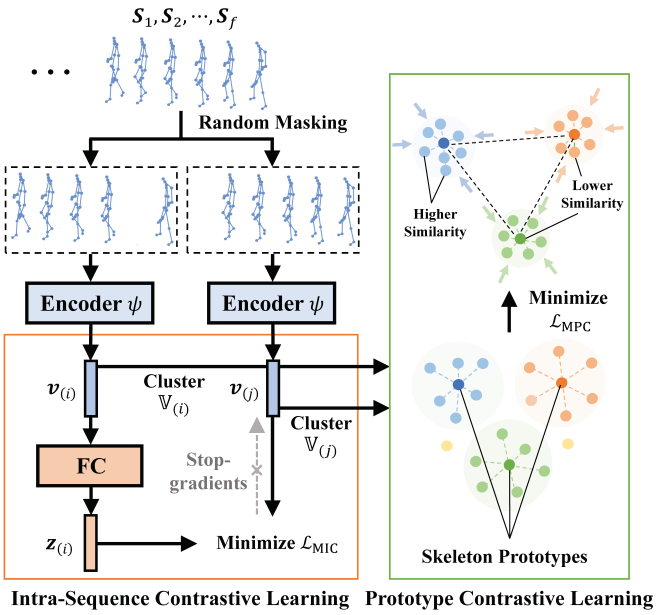


Figure 2: Schematics of our framework with masked prototype contrastive learning and masked intra-sequence contrastive learning.

any label, such that the representation $\boldsymbol{v}_i^{\mathcal{P}}$ in probe set can match the representation $\boldsymbol{v}_j^{\mathcal{G}}$ of the same identity in gallery set. The overview of our framework is presented in Fig. 2.

As shown in Fig. 2, we firstly randomly mask each input skeleton sequence to sample i^{th} and j^{th} subsequences, which are encoded into skeleton instances $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$ (see Sec. 3.1). Secondly, we cluster corresponding instance sets $\mathbb{V}_{(i)}$ and $\mathbb{V}_{(j)}$ individually to generate skeleton prototypes, and then enhance the similarity between instances of same prototype while maximizing the dissimilarity between different ones by minimizing $\mathcal{L}_{\text{MPC}}$. Meanwhile, a Siamese architecture is exploited to learn inherent intra-sequence similarity between $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$ by minimizing $\mathcal{L}_{\text{MIC}}$ (see Sec. 3.2).

3.1 Masked Prototype Contrastive Learning

Each person's skeletons typically possess unique features (*e.g.*, anthropometric attributes), while their corresponding sequences could carry recognizable and highly consistent walking patterns [Murray *et al.*, 1964]. Naturally, we expect the model to exploit the most representative skeleton patterns and traits *within each sequence* for person re-ID. A naïve solution is to cluster skeleton sequences to learn the representative features by direct inter-sequence contrastive learning, while it could overlook some valuable *intra-sequence* representations (*e.g.*, subsequences) that might contain key patterns. To encourage the model to fully mine intra-sequence skeleton features and high-level semantics (*e.g.*, identity-related patterns) from skeleton sequences, we propose a ***masked prototype contrastive learning (MPC) scheme*** to *jointly* focus on the most typical features (***skeleton prototypes***) of different subsequence representations (***skeleton instances***) randomly masked from original sequences, and exploit the instance-prototype similarity and dissimilarity to learn discriminative skeleton representations.

Given an input skeleton sequence $\boldsymbol{S}_{1:f} = (\boldsymbol{S}_1, \cdots, \boldsymbol{S}_f)$, we exploit an MLP encoder with one hidden layer to encode each skeleton as:

$$\boldsymbol{h}_j = \psi(\boldsymbol{S}_j) = \mathbf{W}^2 \sigma\left(\mathbf{W}^1 \boldsymbol{S}_j\right), \tag{1}$$

where $\psi(\cdot)$ represents the encoder function, $\mathbf{W}^1 \in \mathbb{R}^{H\times K}$ and $\mathbf{W}^2 \in \mathbb{R}^{H\times H}$ denote the learnable weight matrices to encode the j^{th} skeleton $\boldsymbol{S}_j \in \mathbb{R}^K$ into a latent feature representation $\boldsymbol{h}_j \in \mathbb{R}^H$, and $\sigma(\cdot)$ is a ReLU non-linear activation function. Then, to sample subsequence representations from the encoded sequence representation $(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f)$ of $\boldsymbol{S}_{1:f}$, we utilize a masking function $\mathcal{M}$ to randomly produce x masks, *i.e.*, zero-masking positions, for each skeleton sequence of length f with:

$$\mathcal{M}(f, x) = (m_1, \cdots, m_f), \tag{2}$$

where $m_j \in \{0, 1\}$ is the mask status for the j^{th} position of a sequence and $\sum_{j=1}^{f} m_j = f - x$. We apply the generated random masks to $\boldsymbol{S}_{1:f}$ and its corresponding skeleton representations $(\boldsymbol{h}_1, \cdots, \boldsymbol{h}_f)$ (see Eq. (1)), which are then integrated into a subsequence representation as (see Fig. 2):

$$\boldsymbol{v}_{(i)} = \frac{1}{f-x} \sum_{j=1}^{f} m_{(i),j} w_j \boldsymbol{h}_j, \tag{3}$$

where $\boldsymbol{v}_{(i)} \in \mathbb{R}^H$ ($i \in \{1, \cdots, q\}$) denotes the feature representation of i^{th} subsequence sampled from $\boldsymbol{S}_{1:f}$ using x random masks, q is the number of subsequence sampling, $m_{(i),j}$ denotes the mask status of the j^{th} position at the i^{th} sampling, while w_j represents the importance of j^{th} skeleton representation $\boldsymbol{h}_j$. Here each skeleton is assumed to equally contribute to representing sequence features, *i.e.*, $w_j = 1$. For clarity, we use $\mathbb{V}_{(i)} = \{\boldsymbol{v}_{(i),j}\}_{j=1}^{N_1}$ to denote all subsequence representations in the i^{th} subsequence sampling of the training set $\Phi_{\mathcal{T}}$. Note that we sample one random subsequence for each training sequence at each sampling. $\mathbb{V}_{(i)} = \{\boldsymbol{v}_{(i),j}\}_{j=1}^{N_1}$ are exploited as *skeleton instances* for the MPC scheme.

To group feature-similar skeleton instances and discover semantic clusters with arbitrary shapes, we leverage the DBSCAN algorithm [Ester *et al.*, 1996] to perform clustering *individually* on the i^{th} instance set $\mathbb{V}_{(i)}$ corresponding to i^{th} subsequence sampling, as shown in Fig. 2, and generate clusters $\overline{\mathbb{V}}_{(i)}^c = \{\boldsymbol{v}_{(i),j}^c\}_{j=1}^{N_c}$, $c \in \{1, \cdots, C\}$, where C is the number of clusters (*i.e.*, pseudo classes), and each cluster $\overline{\mathbb{V}}_{(i)}^c$ contains N_c instances belonging to the c^{th} pseudo class. We *averagely aggregate* instance features of the same cluster to generate the corresponding skeleton prototype as:

$$\boldsymbol{p}_{(i)}^c = \frac{1}{N_c} \sum_{j=1}^{N_c} \boldsymbol{v}_{(i),j}^c, \tag{4}$$

where $\boldsymbol{p}_{(i)}^c \in \mathbb{R}^H$ denotes the skeleton prototype of the c^{th} cluster $\overline{\mathbb{V}}_{(i)}^c$. To jointly focus on the representative skeleton features in all instance sets and encourage capturing high-level skeleton semantics from different prototypes, we exploit

a masked prototype contrastive (MPC) loss to enhance the similarity of each skeleton instance to the corresponding prototype and maximize its dissimilarity to other prototypes by:

$$\mathcal{L}_{\text{MPC}} = \frac{1}{N}\sum_{i=1}^{q}\sum_{c=1}^{C_i}\sum_{j=1}^{N_c} -\log\frac{\exp\left(\boldsymbol{v}_{(i),j}^{c}\cdot\boldsymbol{p}_{(i)}^{c}/\tau\right)}{\sum_{k=1}^{C_i}\exp\left(\boldsymbol{v}_{(i),j}^{c}\cdot\boldsymbol{p}_{(i)}^{k}/\tau\right)}, \tag{5}$$

where N represents the number of all skeleton instances, C_i denotes the number of skeleton prototypes generated from the i^{th} instance set $\mathbb{V}_{(i)}$, N_c is the number of instances belonging to the c^{th} prototype $\boldsymbol{p}_{(i)}^{c}$ in $\mathbb{V}_{(i)}$, and τ represents the temperature for contrastive learning. It is worth noting that the naïve prototype contrastive learning (denoted as NPC) using original sequences is a special case of the proposed MPC scheme when $q = 1$ and $x = 0$ (see Eq. (2) and (3)). The MPC scheme can be viewed as to perform finer prototype learning with different subsequences, and allow the model to jointly attend to key skeleton patterns from different representation subspaces of the original sequences, which encourages capturing more discriminative skeleton features for person re-ID (see Sec. 5). The objective of MPC can be theoretically formulated in the form of Expectation-Maximization (EM) algorithms. We prove the effectiveness of MPC and show its relations to existing contrastive losses in Appendix A.

3.2 Masked Intra-Sequence Contrastive Learning

The continuity of human motion typically results in very little variation of poses/skeletons within a small temporal interval [Rao *et al.*, 2021b]. Due to this nature, subsequences of the same skeleton sequence usually possess strong inherent correlations. For example, they could locally share similar skeletons and partial sequences with consistent walking patterns. To exploit such intra-sequence relationships and inherent consistency ($e.g.$, pattern invariance) within sequences to learn better skeleton representations, we propose the ***masked intra-sequence contrastive learning (MIC)*** below.

Given two skeleton instances ($i.e.$, subsequence representations), $\boldsymbol{v}_{(i)}$ and $\boldsymbol{v}_{(j)}$, of the same sequence, we first map them into a contrasting space $\mathbb{R}^H$ with a fully-connected (FC) layer $\mathcal{F}_c(\cdot)$ by: $\mathcal{F}_c\left(\boldsymbol{v}_{(i)}\right) = \boldsymbol{z}_{(i)}$ and $\mathcal{F}_c\left(\boldsymbol{v}_{(j)}\right) = \boldsymbol{z}_{(j)}$, where $\boldsymbol{z}_{(i)}, \boldsymbol{z}_{(j)} \in \mathbb{R}^H$. Inspired by [Chen and He, 2021], we leverage a Siamese architecture to contrast one instance in the original feature space with the other one in the new contrasting space, so as to *symmetrically* learn their inherent similarity. To this end, we exploit a masked intra-sequence contrastive learning (MIC) loss to minimize the negative cosine similarity between two instances of the same sequence by:

$$\mathcal{L}_{\text{MIC}} = -\alpha\frac{\boldsymbol{z}_{(i)}}{\left\|\boldsymbol{z}_{(i)}\right\|_2}\cdot\frac{\boldsymbol{v}_{(j)}}{\left\|\boldsymbol{v}_{(j)}\right\|_2} - \beta\frac{\boldsymbol{z}_{(j)}}{\left\|\boldsymbol{z}_{(j)}\right\|_2}\cdot\frac{\boldsymbol{v}_{(i)}}{\left\|\boldsymbol{v}_{(i)}\right\|_2}, \tag{6}$$

where $\|\cdot\|_2$ denotes ℓ_2-norm, α and β are weights for contrastive learning of representation pairs $(\boldsymbol{z}_{(i)}, \boldsymbol{v}_{(j)})$ and $(\boldsymbol{z}_{(j)}, \boldsymbol{v}_{(i)})$, respectively. Here $\mathcal{L}_{\text{MIC}}$ is defined for two subsequence representations of a skeleton sequence and the total loss is averaged over all sequences. To enable more stable and better contrastive learning, we employ a symmetrized loss with equal weights for two contrastive representation pairs, $i.e.$, $\alpha = \beta = 0.5$, and adopt an alternating stop-gradient operation following [Chen and He, 2021] when contrasting each pair, as shown in Fig. 2 (Note that we only visualize one contrastive pair for conciseness). We provide hypotheses and proof for the effectiveness of MIC in Appendix A.

3.3 The Entire Framework

The proposed SimMC framework combines both MPC loss (see Eq. (5)) and MIC loss (see Eq. (6)) to perform unsupervised contrastive learning of skeleton representations with:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{MIC}} + (1-\lambda)\mathcal{L}_{\text{MPC}}, \tag{7}$$

where λ is the weight coefficient to trade off the importance of different contrastive learning. For convenience, here we use $\mathcal{L}_{\text{MIC}}$ to denote the total MIC loss averaging over all training skeleton sequences. To facilitate training and generate more reliable clusters, we optimize our model by alternating clustering and contrastive representation learning. For the person re-ID task, we exploit the encoder $\psi(\cdot)$ learned by our framework to encode each skeleton sequence of the probe set $\Phi_{\mathcal{P}}$ into corresponding representations, $\{\boldsymbol{v}_i^{\mathcal{P}}\}_{i=1}^{N_2}$, which are matched with the representations, $\{\boldsymbol{v}_j^{\mathcal{G}}\}_{j=1}^{N_3}$, of the same identity in the gallery set $\Phi_{\mathcal{G}}$ based on the Euclidean distance.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our framework on four person re-ID benchmark datasets with 3D skeleton data, namely *IAS-Lab* [Munaro *et al.*, 2014b], *KS20* [Nambiar *et al.*, 2017], *BIWI* [Munaro *et al.*, 2014a], *KGBD* [Andersson and Araujo, 2015], and a large-scale multi-view gait dataset *CASIA-B* [Yu *et al.*, 2006], which contain 11, 20, 50, 164, and 124 different individuals, respectively. For BIWI and IAS-Lab, we set each testing set as the gallery and the other one as the probe. For KS20, we randomly take one skeleton sequence from each view as the probe sequence and use one half of the remaining sequences for training and the other half as the gallery. For KGBD, we randomly choose one skeleton video of each individual as the probe set, and equally divide the remaining videos into the training set and gallery set. In CASIA-B, all testing sequences are grouped by three conditions (Normal, Bags, Clothes), and we evaluate our framework with single-condition and cross-condition settings following [Liu *et al.*, 2015]. We repeat experiments with each setup for multiple times and report the average performance.

Implementation Details. We set sequence length f to 6 on IAS-Lab, KS20, BIWI, and KGBD datasets for a fair comparison with existing methods, and empirically employ $x = 2$ random masks for subsequence sampling. For the largest dataset CASIA-B with roughly estimated skeleton data from RGB videos, we set $f = 40$ with $x = 10$ random masks. The number of random subsequence sampling is $q = 2$ and the embedding size for skeleton representations is $H = 256$ for all datasets. We empirically set the temperature $\tau = 0.06$ (KGBD), $\tau = 0.07$ (BIWI), $\tau = 0.075$ (CASIA-B), $\tau = 0.08$ (KS20, IAS-Lab) for MPC learning, and adopt the weight coefficient $\lambda = 0.5$ for KS20, KGBD, and IAS-B, $\lambda = 0.75$

				KS20				KGBD				IAS-A			
Types	**Methods**	**# Params**	**GFLOPs**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
Hand-crafted	D_{13} [Munaro *et al.*, 2014a]	—	—	39.4	71.7	81.7	18.9	17.0	34.4	44.2	1.9	40.0	58.7	67.6	24.5
	D_{16} [Pala *et al.*, 2019]	—	—	51.7	77.1	86.9	24.0	31.2	50.9	59.8	4.0	42.7	62.9	70.7	25.2
Supervised	PoseGait [Liao *et al.*, 2020]	8.93M	121.60	49.4	80.9	90.2	23.5	50.6	67.0	72.6	13.9	28.4	55.7	69.2	17.5
	SGELA [Rao *et al.*, 2021b] + DF	9.09M	7.48	49.7	67.0	77.1	22.2	43.7	58.7	65.0	7.1	18.0	32.1	46.2	13.5
	MG-SCR [Rao *et al.*, 2021d]	0.35M	6.60	46.3	75.4	84.0	10.4	44.0	58.7	64.6	6.9	36.4	59.6	69.5	14.1
	SM-SGE [Rao *et al.*, 2021a] + DF	6.25M	23.92	49.8	78.1	85.2	11.7	43.2	58.6	64.6	7.5	38.5	63.2	73.9	15.0
Self-supervised /Unsupervised	AGE [Rao *et al.*, 2020]	7.15M	37.37	43.2	70.1	80.0	8.9	2.9	5.6	7.5	0.9	31.1	54.8	67.4	13.4
	SGELA [Rao *et al.*, 2021b]	8.47M	7.47	45.0	65.0	75.1	21.2	38.1	53.5	60.0	4.5	16.7	30.2	44.0	13.2
	SM-SGE [Rao *et al.*, 2021a]	5.58M	22.61	45.9	71.9	81.2	9.5	38.2	54.2	60.7	4.4	34.0	60.5	71.6	13.6
	SimMC (Ours)	0.15M	0.99	**66.4**	**80.7**	**87.0**	**22.3**	**54.9**	**66.2**	**70.6**	**11.7**	**44.8**	**65.3**	**72.9**	**18.7**
Unsupervised Fine-tuinig	SGELA + SimMC	8.80M	10.10	*47.3*	*69.7*	*79.3*	20.1	*51.7*	*62.7*	*67.9*	*15.1*	*16.8*	*33.3*	*48.7*	12.0
	MG-SCR + SimMC	0.53M	7.88	*71.1*	*83.6*	*89.1*	*22.7*	*47.4*	*59.3*	*64.9*	*11.0*	*47.2*	*69.0*	*77.3*	*22.4*
	SM-SGE + SimMC	5.89M	25.10	*67.2*	*82.2*	*88.5*	*23.0*	*47.1*	*59.2*	*64.9*	*10.8*	*51.3*	*69.9*	*75.6*	*27.3*

Table 1: Performance comparison with existing state-of-the-art skeleton-based methods on KS20, KGBD, and IAS-A. The amount of network parameters (million (M)) and computational complexity (giga floating-point operations (GFLOPs)) for the deep learning based methods are reported. "+ DF" denotes direct supervised fine-tuning. **Bold** refers to the best cases among self-supervised/unsupervised methods, while *italics* indicate achieving higher performance when exploiting SimMC ("+ SimMC") to fine-tune corresponding pre-trained representations.

		IAS-B				BIWI-W				BIWI-S			
Types	**Methods**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
Hand-crafted	D_{13} [Munaro *et al.*, 2014a]	43.7	68.6	76.7	23.7	14.2	20.6	23.7	17.2	28.3	53.1	65.9	13.1
	D_{16} [Pala *et al.*, 2019]	44.5	69.1	80.2	24.5	17.0	25.3	29.6	18.8	32.6	55.7	68.3	16.7
Supervised	PoseGait [Liao *et al.*, 2020]	28.9	51.6	62.9	20.8	8.8	23.0	31.2	11.1	14.0	40.7	56.7	9.9
	SGELA [Rao *et al.*, 2021b] + DF	23.6	42.9	51.9	14.8	13.9	15.3	16.7	22.9	29.2	65.2	73.8	23.5
	MG-SCR [Rao *et al.*, 2021d]	32.4	56.5	69.4	12.9	10.8	20.3	29.4	11.9	20.1	46.9	64.1	7.6
	SM-SGE [Rao *et al.*, 2021a] + DF	44.3	68.2	77.5	14.9	16.7	31.0	40.2	18.7	34.8	60.6	71.5	12.8
Self-supervised /Unsupervised	AGE [Rao *et al.*, 2020]	31.1	52.3	64.2	12.8	11.7	21.4	27.3	12.6	25.1	43.1	61.6	8.9
	SGELA [Rao *et al.*, 2021b]	22.2	40.8	50.2	14.0	11.7	14.0	14.7	19.0	25.8	51.8	64.4	**15.1**
	SM-SGE [Rao *et al.*, 2021a]	38.9	64.1	75.8	13.3	13.2	25.8	33.5	15.2	31.3	56.3	69.1	10.1
	SimMC (Ours)	**46.3**	**68.1**	**77.0**	**22.9**	**24.5**	**36.7**	**44.5**	**19.9**	**41.7**	**66.6**	**76.8**	12.3
Unsupervised Fine-tuinig	SGELA + SimMC	21.2	39.1	48.8	14.0	*18.4*	*23.1*	*25.0*	*28.7*	*51.8*	*71.3*	*74.4*	*43.3*
	MG-SCR + SimMC	*52.4*	*72.0*	*78.8*	*29.1*	*25.1*	*37.5*	*46.4*	*20.3*	*28.3*	*51.6*	*64.8*	*10.9*
	SM-SGE + SimMC	*55.3*	*72.6*	*80.3*	*34.1*	*25.9*	*39.2*	*45.2*	*22.4*	*42.6*	*64.8*	*76.2*	*15.4*

Table 2: Performance comparison on IAS-B, BIWI-Walking (BIWI-W), and BIWI-Still (BIWI-S). **Bold** refers to the best cases among self-supervised/unsupervised methods, while *italics* indicate achieving higher performance with the fine-tuning of SimMC.

for IAS-A, and $\lambda = 0.25$ for BIWI and CASIA-B. We employ Adam optimizer with learning rate 0.00035 and batch size 256 for all datasets. To perform unsupervised fine-tuning with SimMC, we train SimMC on the unlabeled skeleton representations pre-trained by original models, and exploit the skeleton representations learned by SimMC for person re-ID. More implementation details are provided in Appendix B.

Evaluation Metrics. We compute Cumulative Matching Characteristics (CMC) curve and adopt top-1/top-5/top-10 accuracy and Mean Average Precision (mAP) [Zheng *et al.*, 2015] to quantitatively evaluate person re-ID performance.

4.2 Comparison with State-of-the-Arts

We compare our framework with existing state-of-the-art self-supervised and unsupervised skeleton-based methods on KS20, KGBD, IAS-Lab, and BIWI in Table 1 and 2. We also include the latest supervised skeleton-based methods and representative hand-crafted methods as a performance reference.

Comparison with Self-supervised and Unsupervised Methods. Our framework shows evident advantages in terms of performance and efficiency over existing state-of-the-art self-supervised and unsupervised methods. As reported in Table 1 and 2, SimMC significantly outperforms AGE [Rao *et al.*, 2020] and SM-SGE [Rao *et al.*, 2021a] that manually design pretext tasks based on pre-defined skeleton modeling such as skeleton graphs by a large margin of 7.4-52.0% top-1 accuracy and 2.2-13.4% mAP on all datasets. Compared with the SGELA model [Rao *et al.*, 2021b] using direct inter-sequence contrastive learning, our framework achieves remarkably better performance on five out of six testing sets (KS20, KGBD, IAS-A, IAS-B, BIWI-W) by up to 28.1% top-1 accuracy and 8.9% mAP, which demonstrates that the proposed SimMC combining both prototype (MPC) and intra-sequence contrastive learning (MIC) can capture more discriminative features within skeleton sequences for person re-ID on different datasets. Notably, SimMC also enjoys the smallest model size (only 0.15M) for skeleton representation learning among all approaches shown in Table 1, which suggests its higher model efficiency for person re-ID tasks.

By applying the proposed framework to fine-tuning SGELA and SM-SGE models, we can further improve their performance with an average gain of 16.9% and 8.1% top-1 accuracy respectively on all datasets. Such results demonstrate both effectiveness and scalability of proposed masked contrastive learning, which is compatible with existing models and can fully exploit their pre-trained features to achieve higher-quality skeleton representations for person re-ID.

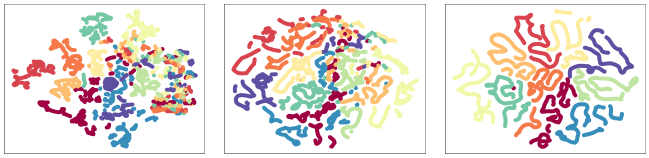
Comparison with Hand-crafted and Supervised Methods. In contrast to hand-crafted methods (D_{13} and D_{16}) that rely on geometric joint distances and anthropometric descriptors, our approach obtains similar performance on IAS testing sets, while it achieves a distinct improvement of 7.5-37.9% top-1 accuracy on BIWI, KS20, and KGBD datasets that contain more views and individuals. Despite utilizing *unlabeled* skeleton data as the sole input, the proposed SimMC still

Probe-Gallery	Normal-Normal				Bags-Bags				Clothes-Clothes				Clothes-Normal				Bags-Normal			
Methods	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**	**top-1**	**top-5**	**top-10**	**mAP**
ELF [Gray and Tao, 2008]	12.3	35.6	50.3	—	5.8	25.5	37.6	—	19.9	43.9	56.7	—	5.6	16.0	26.3	—	17.1	30.0	37.9	—
SDALF [Farenzena *et al.*, 2010]	4.9	27.0	41.6	—	10.2	33.5	47.2	—	16.7	42.0	56.7	—	11.6	19.4	27.6	—	22.9	30.1	36.1	—
MLR [Liu *et al.*, 2015]	16.3	43.4	60.8	—	18.9	44.8	59.4	—	25.4	53.3	68.9	—	20.3	42.6	**56.9**	—	31.8	53.6	64.1	—
AGE [Rao *et al.*, 2020]	20.8	29.3	34.2	3.5	37.1	56.2	67.0	9.8	35.5	54.3	65.3	9.6	14.6	33.0	42.7	3.0	32.4	51.2	60.1	3.9
SM-SGE [Rao *et al.*, 2021a]	50.2	73.5	81.9	6.6	26.6	49.0	59.4	9.3	27.2	51.4	63.2	9.7	10.6	26.3	35.9	3.0	16.6	36.8	47.5	3.5
SGELA [Rao *et al.*, 2021b]	71.8	87.5	91.4	9.8	48.1	69.5	77.7	**16.5**	51.2	73.8	81.5	7.1	15.9	30.8	40.6	4.7	36.4	57.1	64.6	6.7
SimMC (Ours)	**84.8**	**92.3**	**93.7**	**10.8**	**69.1**	**86.6**	**91.3**	**16.5**	**68.0**	**88.1**	**93.0**	**15.7**	**25.6**	**43.8**	54.0	**5.4**	**42.0**	**59.8**	**68.9**	**7.1**

Table 3: Comparison with appearance-based and skeleton-based methods on CASIA-B. "Bags-Normal" represents the probe set with "Bags" condition and gallery set with "Normal" condition. "—" indicates no published result. Full results are in Appendix B.

	IAS-A		IAS-B		BIWI-S		BIWI-W		KS20		KGBD	
Configurations	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**	**top-1**	**mAP**
Baseline	29.4	13.8	30.2	13.3	24.8	9.3	10.9	14.1	17.0	9.5	34.5	6.4
NPC	39.2	17.8	40.7	21.5	38.1	11.3	21.2	18.3	64.8	20.5	53.0	11.0
MPC	43.1	18.5	43.8	22.3	40.1	11.7	23.7	19.5	65.6	21.1	53.6	11.0
MPC + MIC	44.8	18.7	46.3	22.9	41.7	12.3	24.5	19.9	66.4	22.3	54.9	11.7

Table 4: Ablation study of framework with different configurations: Naïve prototype contrastive learning (NPC) using only original sequences, masked prototype contrastive learning (MPC) scheme and corresponding masked intra-sequence contrastive learning (MIC).



(a) AGE (b) SM-SGE (c) SimMC

Figure 3: t-SNE visualization of representations learned by AGE (a), SM-SGE (b), and SimMC (c) for first ten classes in BIWI. Different colors denote skeleton representations of different classes.

performs better than the latest supervised models PoseGait and MG-SCR in most cases. Interestingly, applying SimMC to SM-SGE achieves significantly higher performance gains than direct supervised fine-tuing (DF) in terms of top-1 accuracy (3.9-17.4%), top-5 accuracy (0.6-6.7%), top-10 accuracy (0.3-5.0%), and mAP (3.3-19.2%) on all datasets. With highly efficient performance and strong scalability, the proposed unsupervised SimMC can be a more general framework for skeleton-based person re-ID and related tasks.

5 Further Analysis

Application to Model-estimated Skeletons. To verify the effectiveness of SimMC when applied to RGB-based scenes with model-estimated 3D skeletons, we utilize pre-trained pose estimation models to extract skeleton data from RGB videos of CASIA-B, and compare the performance of SimMC with representative appearance-based and skeleton-based methods. As shown in Table 3, the proposed SimMC remarkably outperforms state-of-the-art skeleton-based models SM-SGE and SGELA by a distinct margin of 5.6% to 42.5% top-1 accuracy and 0.4% to 8.6% mAP in different conditions, which suggests the stronger ability of our framework on capturing discriminative features from estimated skeleton data for person re-ID. Compared with appearance-based ELF and MLR models that utilize visual features (*e.g.*, colors, textures, and silhouettes) with extra label information, the skeleton-based SimMC also achieves superior performance in all conditions of CASIA-B, which demonstrates its great applicable

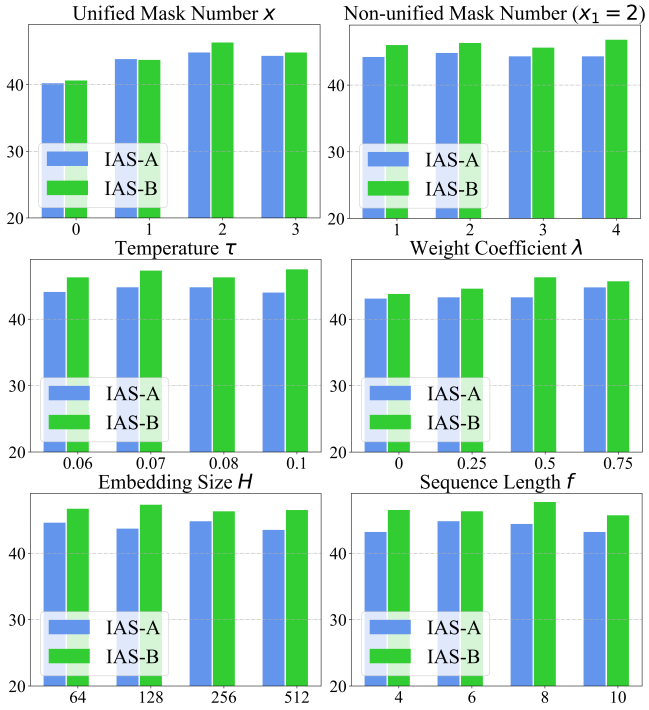


Figure 4: Top-1 accuracy on IAS-A/B showing effects of hyper-parameters. "Non-unified Mask Number ($x_1 = 2$)" denotes using different mask numbers including $x = 2$ for subsequence sampling.

value and potential for person re-ID under large-scale RGB-based scenarios and more general settings.

Ablation Study. We conduct ablation study to demonstrate the contribution of each component in our framework. We adopt 3D coordinates of raw skeleton sequences as the baseline representation for person re-ID. As reported in Table 4, the model exploiting NPC significantly outperforms the baseline by 9.8-47.8% top-1 accuracy and 2.0-11.0% mAP. Considering that NPC is a special case of the proposed MPC scheme (see Sec. 3.1), such results verify the effectiveness of the skeleton prototype contrastive learning in MPC, which

can capture highly discriminative features within unlabeled skeleton sequences for the task of person re-ID. Employing the standard MPC scheme with randomly sampled subsequences consistently improves the model performance by up to 3.9% top-1 accuracy and 1.2% mAP on all datasets, which demonstrates that MPC is able to mine more representative key features from skeleton subsequences to perform person re-ID. Finally, incorporating MIC into MPC further improves model performance with 0.8-2.5% top-1 accuracy and 0.2%-1.2% mAP gains on different datasets. This justifies our claim that capturing inherent intra-sequence similarity and pattern consistency within sequences could facilitate learning better representations of skeleton sequences for person re-ID.

Discussions. As shown in Fig. 3, we conduct a t-SNE visualization [Van der Maaten and Hinton, 2008] of representations. The skeleton representations learned by our framework are clustered with higher inter-class separation than AGE and SM-SGE, which suggests that SimMC may learn richer class-related semantics and lower-entropy skeleton representations. We also show effects of different parameters on SimMC in Fig. 4, which indicates that the use of random masks ($x > 0$) is the key to the proposed masked contrastive learning, regardless of adopting unified or non-unified mask numbers, while an appropriate fusion ($\lambda > 0$) of MIC and MPC facilitates better skeleton representation learning for person re-ID. Our framework with the optimal parameter setting is not sensitive to changes of some parameters such as temperatures τ. More results and proof are provided in the appendices.

6 Conclusion

In this paper, we propose a simple masked contrastive learning (SimMC) framework to efficiently learn representations of unlabeled skeleton sequences for unsupervised person re-ID. A novel masked prototype contrastive learning (MPC) scheme is devised to cluster the most typical skeleton features of subsequences randomly masked from original sequences, so as to contrast their inherent similarity to learn a discriminative skeleton representation from unlabeled skeletons. To fully exploit inherent relationships between subsequences, we propose a masked intra-sequence contrastive learning (MIC) to learn their similarity and pattern consistency within the sequence for more effective skeleton representations. Our framework outperforms existing state-of-the-art skeleton-based methods and also enjoys high scalability and efficiency to be applied to different models and scenes.

Ethics Statement

Person re-ID as an important emerging research topic possesses great value for both academia and industry. However, illegal or improper use of person re-ID technologies could pose serious threat to the public privacy and society security. Therefore, it should be noted that all datasets used in our experiments are officially shared by reliable public (IAS-Lab, BIWI, KGBD) or private research agency (KS20, CASIA-B), which have guaranteed that the collecting, processing, releasing, and using of all data are with the consent of participated subjects. For the protection of privacy, all individuals are anonymized with simple identity numbers. Our models and codes must only be used for the purpose of research.

Acknowledgements

This work was supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore.

References

[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.

[Chen and He, 2021] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[Farenzena *et al.*, 2010] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, pages 2360–2367. IEEE, 2010.

[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.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020.

[Li *et al.*, 2021] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

[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.

[Liu *et al.*, 2015] Zheng Liu, Zhaoxiang Zhang, Qiang Wu, and Yunhong Wang. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144–1156, 2015.

[Munaro *et al.*, 2014a] 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.*, 2014b] 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. *Journal of Bone and Joint Surgery*, 46(2):335–360, 1964.

[Nambiar *et al.*, 2017] Athira Nambiar, Alexandre Bernardino, Jacinto C Nascimento, and Ana Fred. Context-aware person re-identification in the wild via fusion of gait and anthropometric features. In *International Conference on Automatic Face & Gesture Recognition*, pages 973–980. IEEE, 2017.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[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*, 2019.

[Rao *et al.*, 2020] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Huang Da, Jun Cheng, and Bin Hu. Self-supervised gait encoding with locality-aware attention for person re-identification. In *IJCAI*, volume 1, pages 898–905, 2020.

[Rao *et al.*, 2021a] Haocong Rao, Xiping Hu, Jun Cheng, and Bin Hu. Sm-sge: A self-supervised multi-scale skeleton graph encoding framework for person re-identification. In *Proceedings of the 29th ACM international conference on Multimedia*, 2021.

[Rao *et al.*, 2021b] Haocong Rao, Siqi Wang, Xiping Hu, Mingkui Tan, Yi Guo, Jun Cheng, Xinwang Liu, and Bin Hu. A self-supervised gait encoding approach with locality-awareness for 3d skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Rao *et al.*, 2021c] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences*, 569:90–109, 2021.

[Rao *et al.*, 2021d] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification. In *IJCAI*, 2021.

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

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018.

[Ye *et al.*, 2021] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[Yu *et al.*, 2006] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444. IEEE, 2006.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.