TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification

Haocong Rao   Chunyan Miao*

LILY Research Center, Nanyang Technological University, Singapore

School of Computer Science and Engineering, Nanyang Technological University, Singapore

{haocong001, ascymiao}@ntu.edu.sg

Abstract

Person re-identification (re-ID) via 3D skeleton data is an emerging topic with prominent advantages. Existing methods usually design skeleton descriptors with raw body joints or perform skeleton sequence representation learning. However, they typically cannot concurrently model different body-component relations, and rarely explore useful semantics from fine-grained representations of body joints. In this paper, we propose a generic Transformer-based Skeleton Graph prototype contrastive learning (TranSG) approach with structure-trajectory prompted reconstruction to fully capture skeletal relations and valuable spatial-temporal semantics from skeleton graphs for person re-ID. Specifically, we first devise the Skeleton Graph Transformer (SGT) to simultaneously learn body and motion relations within skeleton graphs, so as to aggregate key correlative node features into graph representations. Then, we propose the Graph Prototype Contrastive learning (GPC) to mine the most typical graph features (graph prototypes) of each identity, and contrast the inherent similarity between graph representations and different prototypes from both skeleton and sequence levels to learn discriminative graph representations. Last, a graph Structure-Trajectory Prompted Reconstruction (STPR) mechanism is proposed to exploit the spatial and temporal contexts of graph nodes to prompt skeleton graph reconstruction, which facilitates capturing more valuable patterns and graph semantics for person re-ID. Empirical evaluations demonstrate that TranSG significantly outperforms existing state-of-the-art methods. We further show its generality under different graph modeling, RGB-estimated skeletons, and unsupervised scenarios. Our codes are available at https://github.com/Kali-Hac/TranSG.
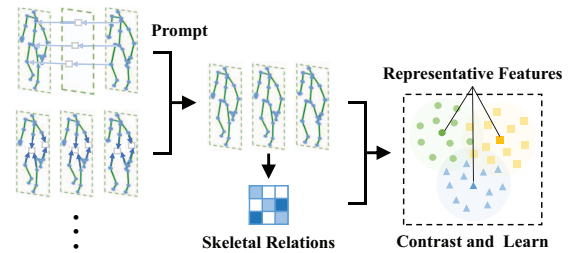
Figure 1. TranSG temporally and spatially masks skeleton graphs to prompt their reconstruction, while integrating full skeletal relations into contrastive learning of typical features for person re-ID.

1. Introduction

Person re-identification (re-ID) is a challenging task of retrieving and matching a specific person across varying views or scenarios, which has empowered many vital applications such as security authentication, human tracking, and robotics [1–4]. Recently driven by economical, non-obtrusive and accurate skeleton-tracking devices like Kinect [5], person re-ID via 3D skeletons has attracted surging attention in both academia and industry [2, 6–14]. Unlike conventional methods that rely on visual appearance features (*e.g.*, colors, silhouettes), skeleton-based person re-ID methods model unique body and motion representations with 3D positions of key body joints, which typically enjoy smaller data sizes, lighter models, and more robust performance under scale and view variations [2, 15].

Most existing methods manually extract anthropometric descriptors and gait attributes from body-joint coordinates [7–10, 13], or leverage sequence learning paradigms such as Long Short-Term Memory (LSTM) [16] to model body and motion features with skeleton sequences [2, 6, 14, 17]. However, these methods rarely explore the inherent body relations within skeletons (*e.g.*, inter-joint motion correlations), thus largely ignoring some valuable skeleton patterns. To fill this gap, a few recent works [11, 18] construct skeleton graphs to model body-component relations

---

*Corresponding author

in terms of structure and action. These methods typically require multi-stage *non-parallel* relation modeling, *e.g.*, [18] models collaborative relations conditioned on the structural relations, while they cannot simultaneously mine different underlying relations. On the other hand, they usually leverage *sequence-level* instances such as averaged features of sequential graphs [11] for representation learning, which inherently limits their ability to exploit richer graph semantics from fine-grained (*e.g.*, node-level) representations. For example, there usually exist strong correlations between nearby body-joint nodes within a local spatial-temporal graph context, which can *prompt* learning unique and recognizable skeleton patterns for person re-ID.

To address the above challenges, we propose a general Transformer-based Skeleton Graph prototype contrastive learning (TranSG) paradigm with structure-trajectory prompted reconstruction as shown in Fig. 1, which integrates different relational features of skeleton graphs and contrasts representative graph features to learn discriminative representations for person re-ID. Specifically, we first model 3D skeletons as graphs, and propose the *Skeleton Graph Transformer (SGT)* to perform *full-relation* learning of body-joint nodes, so as to simultaneously aggregate key relational features of body structure and motion into effective graph representations. Second, a *Graph Prototype Contrastive learning (GPC)* approach is proposed to contrast and learn the most representative skeleton graph features (defined as *"graph prototypes"*) of each identity. By pulling together both *sequence-level* and *skeleton-level* graph representations to their corresponding prototypes while pushing apart representations of different prototypes, we encourage the model to capture discriminative graph features and high-level skeleton semantics (*e.g.*, identity-associated patterns) for person re-ID. Last, motivated by the inherent structural correlations and pattern continuity of body joints [2], we devise a *graph Structure-Trajectory Prompted Reconstruction (STPR) mechanism* to exploit the spatial-temporal context (*i.e.*, graph structure and trajectory) of skeleton graphs to prompt the skeleton graph reconstruction, which facilitates learning richer graph semantics and more effective graph representations for the person re-ID task.

Our key contributions can be summarized as follows:

- We present a generic TranSG paradigm to learn effective representations from skeleton graphs for person re-ID. To the best of our knowledge, TranSG is the first *transformer* paradigm that unifies skeletal relation learning and skeleton graph contrastive learning specifically for skeleton-based person re-ID.

- We devise a skeleton graph transformer (SGT) to fully capture relations in skeleton graphs and integrate key correlative node features into graph representations.

- We propose the graph prototype contrastive learn-

ing (GPC) to contrast and learn the most representative graph features and identity-related semantics from both skeleton and sequence levels for person re-ID.

- We devise the graph structure-trajectory prompted reconstruction (STPR) to exploit spatial-temporal graph contexts for reconstruction, so as to capture more key graph semantics and unique features for person re-ID.

Extensive experiments on five public benchmarks show that TranSG prominently outperforms existing state-of-the-art methods and is highly scalable to be applied to different graph modeling, RGB-estimated or unlabeled skeleton data.

2. Related Works

Person Re-identification Using Skeleton Data. Early methods manually design skeleton descriptors (*e.g.*, anthropometric and gait attributes) from raw body joints. In [13], seven Euclidean distances between certain joints are computed to perform distance metric learning for person re-ID. An further extension with 13 (D_{13}) and 16 skeleton descriptors (D_{16}) in [12] and [7] are utilized to identify different persons. Recent methods employ deep neural networks to perform representation learning with skeleton sequences or graphs. In [8], a CNN-based model PoseGait is devised to encode body-joint sequences and hand-crafted pose features for human recognition. [6] proposes the AGE model to encode recognizable gait features from 3D skeleton sequences, while its extension SGELA [2] further combines self-supervised semantics learning (*e.g.*, sequence sorting) and sequence contrastive learning to improve discriminative feature learning. The SimMC framework [14] is proposed to encode prototypes and intra-sequence relations of masked skeleton sequences for person re-ID. MG-SCR [18] and SM-SGE [19] perform multi-stage body-component relation learning based on multi-scale graphs to learn person re-ID representations. [20] proposes a general skeleton feature re-ranking mechanism for skeleton-based person re-ID.

Contrastive Learning. The aim of contrastive learning is to pull closer homogeneous or positive representation pairs while pushing farther negative pairs in a certain feature space. It has been broadly applied to various areas to learn effective feature representations [2, 11, 14, 21–26]. An instance discrimination paradigm with exemplar tasks is designed in [21] for visual contrastive learning. A contrastive predictive coding (CPC) approach based on the context auto-encoding and probabilistic contrastive loss [27] is proposed to learn different-domain representations. PCL [22] combines k-means clustering and contrastive learning for image representation learning. Skeleton contrastive paradigms based on consecutive sequences or randomly masked sequences are devised in [2] and [14] for unsupervised skeleton representation learning and person re-ID.
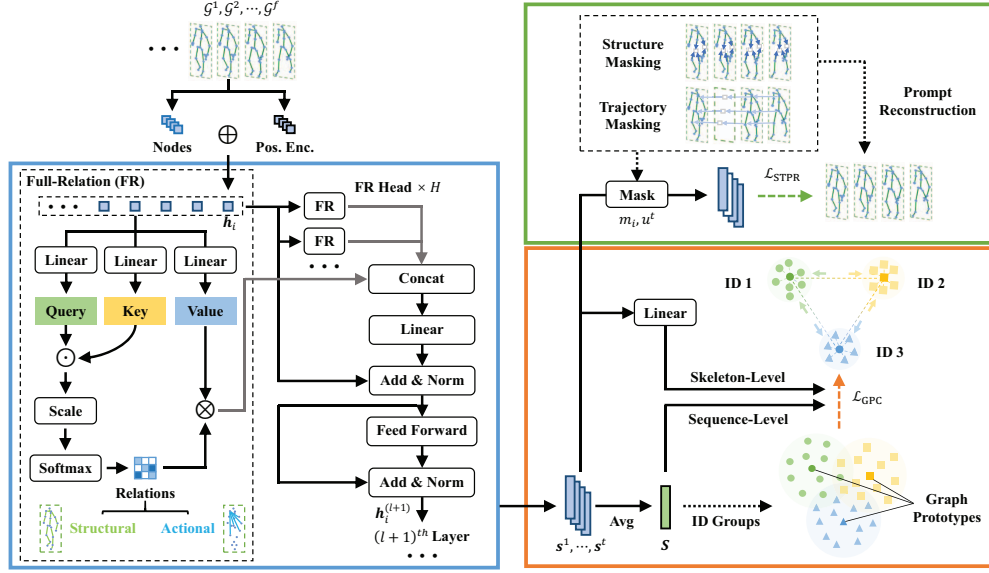
Figure 2. Schematics of our approach with the skeleton graph transformer (SGT) (shown in the blue box), graph prototype contrastive learning (GPC) (presented in the orange box), and graph structure-trajectory prompted reconstruction (STPR) (presented in the green box).

Prompt Learning. The function of prompts is to provide additional knowledge, instruction, or context for the input of models, such that they can be prompted to give more reliable outputs for different tasks [28–36]. In [30], CLIP leverages language-based prompts to generalize the pre-trained visual representations to many tasks. CoOp [31] is further devised to automatically model task-relevant prompts with continuous representations to improve the downstream task performance. As far as we know, this work for the first time explores *graph prompts* (defined as structure and trajectory contexts) for skeleton graph reconstruction, so as to encourage capturing more key features and graph semantics (*e.g.*, pattern continuity) for person re-ID.

3. The Proposed Approach

Given a 3D skeleton sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_f) \in \mathbb{R}^{f \times J \times 3}$, where $\boldsymbol{x}_t \in \mathbb{R}^{J \times 3}$ denotes the tth skeleton with 3D coordinates of J body joints. Each skeleton sequence $\boldsymbol{X}$ corresponds to a person identity y, where y $\in \{1, \cdots, C\}$ and C is the number of different classes (*i.e.*, identities). We use $\Phi_T = \left\{\boldsymbol{X}_i^T\right\}_{i=1}^{N_1}$, $\Phi_P = \left\{\boldsymbol{X}_i^P\right\}_{i=1}^{N_2}$, and $\Phi_G = \left\{\boldsymbol{X}_i^G\right\}_{i=1}^{N_3}$ to denote the *training* set, *probe* set, and *gallery* set that contain N_1, N_2, and N_3 skeleton sequences of different persons in different scenes and views. Our aim is to learn an encoder to map skeleton sequences into effective representations, such that the encoded skeleton sequence representations (denoted as $\{\boldsymbol{S}_i^P\}_{i=1}^{N_2}$) in the probe set can be matched with the representations (denoted as $\{\boldsymbol{S}_i^G\}_{i=1}^{N_3}$) of the same identity in the gallery set.

The overview of our TranSG approach is shown in Fig. 2: Firstly, we represent each skeleton sequence $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_f$ as skeleton graphs $\mathcal{G}^1, \cdots, \mathcal{G}^f$ (see Sec. 3.1), and integrate the graph positional information into their node representations $\boldsymbol{h}_i$. Then, they are fed into the skeleton graph transformer (SGT) to fully capture body-component relations with multiple full-relation (FR) heads (see Sec. 3.2). We employ multiple SGT layers and take the lth layer output $\boldsymbol{h}_i^{(l+1)}$ as the input of $(l+1)^{th}$ layer. Next, the centroids of graph features belonging to different identities (ID) are utilized to generate *graph prototypes*, and we enhance the similarity of both skeleton-level ($\boldsymbol{s}^t$) and sequence-level graph representations ($\boldsymbol{S}$) to their corresponding prototypes, while maximizing their dissimilarity to other prototypes by optimizing $\mathcal{L}_{\text{GPC}}$ (see Sec. 3.3). Meanwhile, we randomly mask skeleton graph structure and node trajectory, and exploit correspondingly masked graph representations as contexts to prompt reconstruction by minimizing $\mathcal{L}_{\text{STPR}}$ (see Sec. 3.4). The skeleton graph representations learned from our approach are exploited for person re-ID (see Sec. 3.5).

3.1. Skeleton Graph Construction

The human body with joints can be naturally modeled as graphs to characterize rich structural and positional information [19]. We construct skeleton graphs with the body joints as nodes and the structural connections between adjacent joints as edges. Each graph $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$ (corresponding to the tth skeleton $\boldsymbol{x}_t$) consists of nodes $\mathcal{V}^t = \{\boldsymbol{v}_1^t, \boldsymbol{v}_2^t, \cdots, \boldsymbol{v}_J^t\}$, $\boldsymbol{v}_i^t \in \mathbb{R}^3$, $i \in \{1, \cdots, J\}$ and edges $\mathcal{E}^t = \{e_{i,j}^t \mid \boldsymbol{v}_i^t, \boldsymbol{v}_j^t \in \mathcal{V}^t\}$, $e_{i,j}^t \in \mathbb{R}$. Here $\mathcal{V}^t$ and $\mathcal{E}^t$ denote the set of nodes corresponding to J different body joints and

the set of their internal connection relations, respectively. The adjacency matrix of $\mathcal{G}^t$ is denoted as $\mathbf{A}^t \in \mathbb{R}^{J \times J}$ to represent the relations among J nodes. $\mathbf{A}^t$ is initialized based on the connections of adjacent body joints.

3.2. Skeleton Graph Transformer

As our goal is to capture discriminative skeleton features for person re-ID, it is crucial to consider two *unique* properties of human skeletons: (1) Body *structural* features, which can be inferred from the relations between adjacent body joints; (2) Skeleton *actional* patterns (*e.g.*, gait [37]), which are typically characterized by the relations among different body components [19]. From the perspective of graphs, we regard each body-joint node as a basic body component, and propose to combine the above relation learning as a *full-relation* learning of body-joint nodes, so as to fully aggregate key body and motion features from skeleton graphs. For this purpose, we devise the *skeleton graph transformer (SGT)* as follows (shown in Fig. 2).

Given a skeleton graph $\mathcal{G}^t(\mathcal{V}^t, \mathcal{E}^t)$ and its adjacency matrix $\mathbf{A}^t$, we first exploit the pre-defined graph structure to generate the positional encoding for graph nodes with:

$$\Delta = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U}, \qquad (1)$$

where $\mathbf{A}, \mathbf{D}$ are the adjacency matrix and degree matrix of the skeleton graph $\mathcal{G}$, respectively, and $\boldsymbol{\Lambda}, \mathbf{U}$ denote the matrices of Laplacian eigenvalues and eigenvectors, respectively. $\mathbf{U}^T\boldsymbol{\Lambda}\mathbf{U}$ is the factorization of the graph Laplacian matrix. For convenience, we use $\mathbf{A}$ to represent $\mathbf{A}^t$ as skeleton graphs in the same dataset share the identical initialized adjacency matrix. We follow [38] to adopt the K smallest non-trivial eigenvectors as the node positional encoding, denoted as $\boldsymbol{\lambda}_i \in \mathbb{R}^K$ for the node $\boldsymbol{v}_i$. They are mapped into feature spaces of the same dimension d with the affine transformation, which are then added by:

$$\boldsymbol{h}_i = (\mathbf{W}_v \boldsymbol{v}_i + \boldsymbol{b}_v) + (\mathbf{W}_p \boldsymbol{\lambda}_i + \boldsymbol{b}_p), \qquad (2)$$

where $\boldsymbol{h}_i \in \mathbb{R}^d$ denotes the ith position-encoded node representation, $\mathbf{W}_v \in \mathbb{R}^{d \times 3}, \mathbf{W}_p \in \mathbb{R}^{d \times K}, \boldsymbol{b}_v, \boldsymbol{b}_p \in \mathbb{R}^d$ are the learnable parameters of the feature transformation for the ith node $\boldsymbol{v}_i$ and its corresponding positional encoding. Intuitively, the addition of positional encoding in Eq. (2) helps preserve the unique positional information of nodes based on the graph structure, *i.e.*, structurally nearby nodes are endowed with similar positional features while the farther nodes possess more dissimilar positional features, so as to encourage more effective node representation learning [38].

Then, given the body-joint node representations, we capture their inherent relations using multiple *independent* full-relation (FR) heads (see Fig. 2), and update node representations by aggregating corresponding relational features:

$$\boldsymbol{w}_{i,j}^{k,l} = \mathrm{Softmax}_j \left(\frac{(\boldsymbol{Q}^{k,l} \boldsymbol{h}_i^{(l)}) \cdot (\boldsymbol{K}^{k,l} \boldsymbol{h}_j^{(l)})}{\sqrt{d_{\mathrm{k}}}} \right), \quad (3)$$

$$\hat{\boldsymbol{h}}_i^{(l)} = \boldsymbol{O}^l \Big\|_{k=1}^{H} \left(\sum_{j \in \mathcal{N}_i} \boldsymbol{w}_{i,j}^{k,l} \boldsymbol{V}^{k,l} \boldsymbol{h}_j^{(l)} \right), \qquad (4)$$

where $\boldsymbol{Q}^{k,l}, \boldsymbol{K}^{k,l}, \boldsymbol{V}^{k,l} \in \mathbb{R}^{d_k \times d}$ are the parameter matrices for query, key, and value transformations in the kth FR head of the lth SGT layer, $\boldsymbol{O}^l \in \mathbb{R}^{d \times d}$ is the parameter matrix for output transformation of the lth SGT layer. $\frac{1}{\sqrt{d_k}}$ is the scaling factor for the scaled dot-product similarity, $\boldsymbol{w}_{i,j}^{k,l}$ denotes the normalized relation value between the ith and j^{th} node captured by the kth FR head in the lth layer, $\|$ represents the concatenation operation, and H is the number of FR heads. For clarity, we use $\hat{\boldsymbol{h}}_i^{(l)} \in \mathbb{R}^d$ denotes the ith node representation that concatenates node features learned from different heads in the lth layer. It is worth noting that SGT naturally generalizes the *self-attention* [39] to the full-relation learning of graph nodes, and can be viewed as a general paradigm that *simultaneously* captures structural and actional relations from both adjacent and non-adjacent body-component nodes. The multiple FR heads enable the model to *jointly* attend to node relations from different feature subspaces and integrate more key correlative node features into final node representations. We follow [38] to apply a Feed Forward Network (FFN) with residual connections [40] and batch normalization [41] by:

$$\overline{\boldsymbol{h}}_i^{(l)} = \mathrm{Norm}\left(\boldsymbol{h}_i^{(l)} + \hat{\boldsymbol{h}}_i^{(l)}\right), \qquad (5)$$

$$\boldsymbol{h}_i^{(l+1)} = \mathrm{Norm}\left(\overline{\boldsymbol{h}}_i^{(l)} + \mathbf{W}_2^l \, \sigma\left(\mathbf{W}_1^l \overline{\boldsymbol{h}}_i^{(l)}\right)\right). \qquad (6)$$

In Eq. (5) and (6), $\mathrm{Norm}(\cdot)$ denotes the batch normalization operation, $\mathbf{W}_1^l \in \mathbb{R}^{2d \times d}, \mathbf{W}_2^l \in \mathbb{R}^{d \times 2d}$ are the learnable parameters of FFN, $\sigma(\cdot)$ is the ReLU activation function, $\overline{\boldsymbol{h}}_i^{(l)}$ and $\boldsymbol{h}_i^{(l+1)}$ represent the intermediate and *output* node representations of the lth SGT layer, respectively. We average the node features in each skeleton graph as the corresponding graph representation, and then integrate f consecutive graph representations into the final sequence-level graph representation $\boldsymbol{S}$ with:

$$\boldsymbol{S} = \frac{1}{f}\sum_{t=1}^{f} \boldsymbol{s}^t = \frac{1}{f}\sum_{t=1}^{f}\frac{1}{J}\sum_{i=1}^{J}\boldsymbol{h}_i^t, \qquad (7)$$

where $\boldsymbol{S}, \boldsymbol{s}^t \in \mathbb{R}^d$ are the *sequence-level* and *skeleton-level* graph representations, corresponding to a skeleton sequence $\boldsymbol{X}$ and the tth skeleton $\boldsymbol{x}_t$ in the sequence, respectively. For simplicity of presentation, we use $\boldsymbol{h}_i^t$ to denote the encoded representation of ith node in the tth skeleton graph. Here we assume that each node representation with aggregated relational features (see Eq. (3), (4)) contributes equally to the graph representation, and each skeleton graph has the same importance in representing patterns of an individual.

3.3. Graph Prototype Contrastive Learning

Each individual's skeletons usually share the same anthropometric features (*e.g.*, skeletal lengths), while their continuous sequence can characterize identity-specific walking patterns (*i.e.*, gait) [14]. In this context, it is desirable to mine the most representative skeleton features (defined as *"prototypes"*) of each individual to learn distinguishable patterns. A straightforward way is to cluster sequence representations to mine prototypes for contrastive learning like [11,14], while they can only generate identity-agnostic (*i.e.*, pseudo-labeled) prototypes with large uncertainty or unreliability, *e.g.*, when existing large intra-class variation, two same-class sequences with diverse patterns might be clustered to two prototype groups. To encourage the model to generate representatives graph features more reliably, we propose to exploit the graph feature centroid of each *ground-truth* identity as *graph prototypes*, which are contrasted with both sequence-level and skeleton-level graph features to learn discriminative representations for person re-ID.

Given the encoded graph representations $\{S_i^T\}_{i=1}^{N_1}$ of training skeleton sequences $\{X_i^T\}_{i=1}^{N_1}$, we group them by ground-truth classes as $\{\widehat{\mathbb{S}}_k\}_{k=1}^C$, where $\widehat{\mathbb{S}}_k = \{S_{k,j}\}_{j=1}^{n_k}$ denotes the set of graph representations belonging to the kth identity, $S_{k,j}$ is the jth sequence-level graph representation, and n_k represents the number of k-class sequences. Then, the graph prototype is generated by averaging the graph features of the same class with:

$$c_k = \frac{1}{n_k} \sum_{j=1}^{n_k} S_{k,j} , \qquad (8)$$

where $c_k \in \mathbb{R}^d$ denotes the graph prototype of the kth identity. To focus on the representative graph prototype of each identity to learn discriminative identity-related semantics from both skeleton and sequence levels, we propose the graph prototype contrastive (GPC) loss as:

$$\mathcal{L}_{\text{GPC}} = \alpha \mathcal{L}_{\text{GPC}}^{seq} + (1 - \alpha) \mathcal{L}_{\text{GPC}}^{ske}, \qquad (9)$$

where

$$\mathcal{L}_{\text{GPC}}^{seq} = \frac{1}{N_1} \sum_{k=1}^C \sum_{j=1}^{n_k} -\log \frac{\exp\left(S_{k,j} \cdot c_k / \tau_1\right)}{\sum_{m=1}^C \exp\left(S_{k,j} \cdot c_m / \tau_1\right)}, \qquad (10)$$

$$\mathcal{L}_{\text{GPC}}^{ske} = \frac{1}{fN_1} \sum_{k=1}^C \sum_{j=1}^{n_k} \sum_{t=1}^f -\log \frac{\exp\left(\mathcal{F}_1\left(s_{k,j}^t\right) \cdot \mathcal{F}_2\left(c_k\right) / \tau_2\right)}{\sum_{m=1}^C \exp\left(\mathcal{F}_1\left(s_{k,j}^t\right) \cdot \mathcal{F}_2\left(c_m\right) / \tau_2\right)}. \qquad (11)$$

In Eq. (9), α is the weight coefficient to fuse sequence-level ($\mathcal{L}_{\text{GPC}}^{seq}$) and skeleton-level graph prototype contrastive learning ($\mathcal{L}_{\text{GPC}}^{ske}$). In Eq. (10) and (11), c_m represents the m-class graph prototype, $s_{k,j}^t$ denotes the graph representation of the tth skeleton in the sequence that corresponds

to $S_{k,j}$ belonging to the kth identity, τ_1, τ_2 are the temperatures for contrastive learning, and $\mathcal{F}_1(\cdot)$, $\mathcal{F}_2(\cdot)$ are linear projection heads to transform skeleton-level graph representations and graph prototypes into the same contrastive space $\mathbb{R}^d$. It should be noted that the graph prototypes are generated from higher level (*i.e.*, sequence-level) representations and the learnable linear projection in Eq. (11) can be viewed as integrating related graph features from both levels for contrastive learning. The proposed GPC loss is essentially a generalized skeleton prototype contrastive loss that combines joint-level relation encoding (see Sec. 3.2), skeleton-level and sequence-level graph learning (see Eq. (9)). Its objective can be theoretically modeled as an Expectation-Maximization (EM) solution (see Appendix I).

3.4. Graph Structure-Trajectory Prompted Reconstruction Mechanism

To exploit more valuable graph features and high-level semantics (*e.g.*, pattern consistency) from both spatial and temporal contexts of skeleton graphs, we propose a self-supervised *graph Structure and Trajectory Prompted Reconstruction (STPR)* mechanism. Motivated by the structural correlations and local motion continuity [2] of body components, we devise two graph context based prompts (see Fig. 2), namely (1) partial node positions of the *same* graph and (2) partial node trajectory among *continuous* graphs, to reconstruct the graph structure and dynamics.

Graph Structure Prompted Reconstruction. Given the tth skeleton graph representation with J encoded nodes $(h_1^t, \cdots, h_J^t)$ encoded by SGT (see Eq. (1)-(6)), we first randomly mask node positions to generate the masked graph representation as:

$$\hat{s}^t = \frac{1}{J - a} \sum_{i=1}^J m_i h_i^t, \qquad (12)$$

where a is the number of masks, m_i is the binary mask value (*i.e.*, 0 for masking and 1 for unmasking) applied on the ith node representation h_i^t, and we have $\sum_{i=1}^J m_i = J - a$. With both spatial positional information (Eq. (1), (2)) and relational features (Eq. (3), (4)) integrated into each node, the masked graph representation $\hat{s}^t \in \mathbb{R}^d$ in Eq. (12) retains the context of graph structure, which is then exploited to prompt the skeleton reconstruction by:

$$\hat{x}^t = f_s(\hat{s}^t), \qquad (13)$$

where $f_s(\cdot)$ is an MLP network with one hidden layer for skeleton reconstruction with the structure prompt, and $\hat{x}^t \in \mathbb{R}^{J \times 3}$ is the predicted skeleton. It is worth noting that Eq. (13) not only utilizes the unmasked node representations to reconstruct their corresponding positions, but also exploits them as the context to predict the masked node positions. For conciseness, we denote the predicted i^{th} training skeleton sequence as $\hat{X}_i = \left(\hat{x}^1, \cdots, \hat{x}^f\right) \in \mathbb{R}^{f \times J \times 3}$.

Table 1. Performance comparison with state-of-the-art skeleton-based methods on different datasets. ♠ denotes skeleton graph based methods, † indicates using hand-crafted descriptors, and ‡ refers to sequence representation learning models. **Bold** denotes the best results.

Methods	BIWI-S				BIWI-W				KS20				IAS-A				IAS-B				KGBD			
	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}	mAP	R_1	R_5	R_{10}
$D_{13}^\dagger$ [12]	13.1	28.3	53.1	65.9	17.2	14.2	20.6	23.7	18.9	39.4	71.7	81.7	24.5	40.0	58.7	67.6	23.7	43.7	68.6	76.7	1.9	17.0	34.4	44.2
$D_{16}^\dagger$ [7]	16.7	32.6	55.7	68.3	18.8	17.0	25.3	29.6	24.0	51.7	77.1	86.9	25.2	42.7	62.9	70.7	24.5	44.5	69.1	80.2	4.0	31.2	50.9	59.8
PoseGait† [8]	9.9	14.0	40.7	56.7	11.1	8.8	23.0	31.2	23.5	49.4	80.9	90.2	17.5	28.4	55.7	69.2	20.8	28.9	51.6	62.9	13.9	50.6	67.0	72.6
AGE‡ [6]	8.9	25.1	43.1	61.6	12.6	11.7	21.4	27.3	8.9	43.2	70.1	80.0	13.4	31.1	54.8	67.4	12.8	31.1	52.3	64.2	0.9	2.9	5.6	7.5
SGELA‡ [2]	15.1	25.8	51.8	64.4	19.0	11.7	14.0	14.7	21.2	45.0	65.0	75.1	13.2	16.7	30.2	44.0	14.0	22.2	40.8	50.2	4.5	38.1	53.5	60.0
MG-SCR♠ [18]	7.6	20.1	46.9	64.1	11.9	10.8	20.3	29.4	10.4	46.3	75.4	84.0	14.1	36.4	59.6	69.5	12.9	32.4	56.5	69.4	6.9	44.0	58.7	64.6
SM-SGE♠ [19]	10.1	31.3	56.3	69.1	15.2	13.2	25.8	33.5	9.5	45.9	71.9	81.2	13.6	34.0	60.5	71.6	13.3	38.9	64.1	75.8	4.4	38.2	54.2	60.7
SPC-MGR♠ [11]	16.0	34.1	57.3	69.8	19.4	18.9	31.5	40.5	21.7	59.0	79.0	86.2	24.2	41.9	66.3	75.6	24.1	43.3	68.4	79.4	6.9	40.8	57.5	65.0
SimMC‡ [14]	12.3	41.7	66.6	76.8	19.9	24.5	36.7	44.5	22.3	66.4	80.7	87.0	18.7	44.8	65.3	72.9	22.9	46.3	68.1	77.0	11.7	54.9	66.2	70.6
TranSG♠ (Ours)	**30.1**	**68.7**	**86.5**	**91.8**	**26.9**	**32.7**	**44.9**	**52.2**	**46.2**	**73.6**	**86.3**	**90.2**	**32.8**	**49.2**	**68.5**	**76.2**	**39.4**	**59.1**	**77.0**	**87.0**	**20.2**	**59.0**	**73.1**	**78.2**

Graph Trajectory Prompted Reconstruction. To encourage the model to capture more unique temporal patterns from skeleton graphs, we propose to reconstruct graph trajectories based on their partial dynamics. In particular, we randomly mask the trajectory of each node with:

$$T_i = \frac{1}{f - b} \sum_{t=1}^{f} u^t h_i^t, \qquad (14)$$

where $T_i \in \mathbb{R}^d$ is the *masked* representation of i-node trajectory (*i.e.*, $h_i^1, \cdots, h_i^f$), b is the number of trajectory masks, u^t is the binary mask value applied to the tth position in the i-node trajectory (*i.e.*, i^{th} node representation of the tth skeleton graph), and we have $\sum_{t=1}^{f} u^t = f - b$. The masked trajectory representation T_i preserves partial graph dynamics of different body-joint nodes, which are then used to prompt the skeleton reconstruction by:

$$\overline{x}_i = f_t(T_i). \qquad (15)$$

In Eq. (15), $\overline{x}_i \in \mathbb{R}^{f \times 3}$ denotes the temporal trajectory of the ith body joints predicted by an MLP network $f_t(\cdot)$ with one hidden layer. Based on the unmasked node representations over the temporal dimension, this reconstruction facilitates capturing key temporal dynamics and semantics (*e.g.*, continuous patterns) to infer the node positions missed on the trajectory. We represent the predicted i^{th} training sequence as $\overline{X}_i \in \mathbb{R}^{f \times J \times 3}$ by transposing the original predicted position matrix $(\overline{x}_1, \cdots, \overline{x}_J) \in \mathbb{R}^{J \times f \times 3}$.

We propose the STPR objective to combine both graph structure and trajectory prompted reconstruction as follows:

$$\mathcal{L}_{\text{STPR}} = \beta \mathcal{L}_{\text{STPR}}^{st} + (1 - \beta)\mathcal{L}_{\text{STPR}}^{tr}, \qquad (16)$$

where

$$\mathcal{L}_{\text{STPR}}^{st} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left\| \hat{X}_i - X_i^T \right\|_1, \qquad (17)$$

$$\mathcal{L}_{\text{STPR}}^{tr} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left\| \overline{X}_i - X_i^T \right\|_1. \qquad (18)$$

In Eq. (16), β is the weight coefficient to fuse structure-prompted ($\mathcal{L}_{\text{STPR}}^{st}$) and trajectory-prompted reconstruction

($\mathcal{L}_{\text{STPR}}^{tr}$). $X_i^T \in \mathbb{R}^{f \times J \times 3}$ denotes the ground-truth positions of ith training skeleton sequence, and $\| \cdot \|_1$ represents the ℓ_1 norm. We employ ℓ_1 reconstruction loss following [19, 42], as it can alleviate gradient explosion of large losses while providing sufficient gradients for the positions with small losses to facilitate more precise reconstruction.

3.5. The Entire Approach

We combine the proposed GPC (Sec. 3.3) and STPR (Sec. 3.4) to perform skeleton representation learning with:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{GPC}} + (1 - \lambda)\mathcal{L}_{\text{STPR}}, \qquad (19)$$

where λ is the weight coefficient to fuse two losses for training. For person re-ID, we leverage the learned SGT to encode skeleton sequences of the probe set Φ_P into sequence-level graph representations, $\{S_i^P\}_{i=1}^{N_2}$ (see Eq. (7)), which are matched with the representations, $\{S_i^G\}_{i=1}^{N_3}$, of the same identity in the gallery set Φ_G based on Euclidean distance.

4. Experiments

4.1. Experimental Setups

Datasets. Our approach is evaluated on four skeleton-based person re-ID benchmark datasets: *IAS* [43], *KS20* [44], *BIWI* [12], *KGBD* [9], which contain 11, 20, 50, and 164 different persons. We also verify the generality of TranSG on the RGB-estimated skeleton data from a large-scale multi-view gait dataset *CASIA-B* [45] with 124 individuals under three conditions (Normal (N), Bags (B), Clothes (C)). We adopt the commonly-used probe and gallery settings following [14, 46] for a fair comparison.

Implementation Details. The number of body joints is $J = 20$ in KGBD, IAS, BIWI, $J = 25$ in KS20, and $J = 14$ in the estimated skeleton data of CASIA-B. The sequence length is set to $f = 6$ for the Kinect-based skeleton datasets (IAS, KS20, BIWI, and KGBD) and $f = 40$ for the RGB-estimated skeleton data (CASIA-B), following existing methods for a fair comparison. The embedding size of each node representation is $d = 128$. We empirically employ 2 SGT layers with $H = 8$ FR heads and $d_k = 16$ for each layer. Each component in our approach is equally

fused with $\alpha = \beta = \lambda = 0.5$. For models trained with RGB-estimated skeletons, we set $\alpha = 1.0$. $\tau_1 = 0.07$ and $\tau_2 = 14$ are empirically used for contrastive learning. The numbers of random masks are set to $a = 10$ and $b = 2$. We use the Adam optimizer with a learning rate of 3.5×10^{-4} for model optimization, and the batch size is set to 256 for all datasets. More details are provided in Appendix II.

Evaluation Metrics. We compute the Cumulative Matching Characteristics (CMC) curve and adopt Rank-1 accuracy (R_1), Rank-5 accuracy (R_5), and Rank-10 accuracy (R_{10}) as performance metrics. Mean Average Precision (mAP) [47] is also used to evaluate the overall performance.

4.2. Comparison with State-of-the Art Methods

We compare our approach with state-of-the-art graph-based methods, hand-crafted methods, and sequence learning methods on BIWI, KS20, IAS, and KGBD in Table 1.

Comparison with Graph-based Methods: As shown in Table 1, the proposed TranSG significantly outperforms two state-of-the-art graph-based methods, MG-SCR [18] and SM-SGE [19], by 11.7-36.7% for mAP and 12.8-48.6% for Rank-1 accuracy on different datasets. As these methods rely on multi-stage relation modeling and sequence-level context learning, the results demonstrate that the proposed full-relation learning model (SGT) with fine-grained (*i.e.*, graph and node level) semantics learning (STPR) is able to learn more unique skeleton features for person re-ID. Compared with the latest SPC-MGR model [11] that utilizes sequence-level graph representations for contrastive learning, our model achieves superior performance with a large margin of 7.5% to 24.5% for mAP and 7.3% to 34.6% for Rank-1 accuracy on all datasets. This suggests the higher effectiveness of our approach combining both sequence-level and skeleton-level graph prototype contrastive learning. We will also show the generality of our model under different-scale skeleton graph modeling in Sec. 5.

Comparison with Hand-crafted and Sequence Learning Methods: In contrast to the methods using hand-crafted anthropometric descriptors (D_{13} [12], D_{16} [7]) or 3D pose features (PoseGait [8]), our approach consistently achieves higher performance by up to 27.3% for mAP and 54.7% for Rank-1 accuracy on all datasets. TranSG also achieves a remarkable improvement over existing sequence representation learning models (AGE [6], SGELA [2], SimMC [14]) in terms of mAP (7.0-37.3%), Rank-1 (4.1-56.1%), Rank-5 (3.2-67.5%), and Rank-10 accuracy (3.2-70.7%). This demonstrates the superiority of our graph-based model, as it can fully capture body relations and discriminative patterns from skeleton graphs for the person re-ID task.

4.3. Ablation Study

We conduct ablation study to verify the effectiveness of each component in our approach. The skeleton sequences

Table 2. Ablation study with different configurations: Naïve prototype contrastive learning (PC), skeleton graph transformer (SGT) with direct supervised learning (DS) or graph prototype contrastive learning (GPC), and structure-trajectory prompted reconstruction (STPR). "+" indicates employing the component.

Configurations	BIWI-S		BIWI-W		KS20		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
Baseline	9.3	24.8	14.1	10.9	9.5	17.0	6.4	34.5
PC	11.3	38.1	18.3	21.2	20.5	64.8	11.0	53.0
SGT + DS	19.0	42.4	21.1	21.7	27.6	60.0	11.1	51.5
SGT + GPC	26.7	66.6	25.5	31.2	42.5	71.3	18.1	57.0
SGT + GPC + STPR	30.1	68.7	26.9	32.7	46.2	73.6	20.2	59.0

Table 3. Performance comparison with *appearance* or *skeleton* based methods on CASIA-B. "B-N" denotes matching the "Bags (B)" probe set with the "Normal (N)" gallery set. "—" indicates no published result. Full results are provided in Appendix II.

	Probe-Gallery	N-N		B-B		C-C		C-N		B-N	
	Methods	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
Appearance	LMNN [48]	—	3.9	—	18.3	—	17.4	—	11.6	—	23.1
	ITML [49]	—	7.5	—	19.5	—	20.1	—	10.3	—	21.8
	ELF [50]	—	12.3	—	5.8	—	19.9	—	5.6	—	17.1
	SDALF [51]	—	4.9	—	10.2	—	16.7	—	11.6	—	22.9
	MLR (Scores) [46]	—	13.6	—	13.6	—	13.5	—	9.7	—	14.7
	MLR (Features) [46]	—	16.3	—	18.9	—	25.4	—	20.3	—	31.8
Skeleton	AGE [6]	3.5	20.8	9.8	37.1	9.6	35.5	3.0	14.6	3.9	32.4
	SM-SGE [19]	6.6	50.2	9.3	26.6	9.7	27.2	3.0	10.6	3.5	16.6
	SPC-MGR [11]	9.1	71.2	11.4	44.3	11.8	48.3	4.3	22.4	4.6	28.9
	SGELA [2]	9.8	71.8	16.5	48.1	7.1	51.2	4.7	15.9	6.7	36.4
	SimMC [14]	10.8	**84.8**	16.5	**69.1**	15.7	**68.0**	5.4	**25.6**	7.1	42.0
	TranSG (Ours)	**13.1**	78.5	**17.9**	67.1	**15.7**	65.6	**6.7**	23.0	**8.6**	**44.1**

of concatenated joints are adopted as the baseline, and we include the naïve prototype contrastive learning (PC) using original sequences [14] for comparison. As shown in Table 2, compared with using raw skeleton sequences or PC without graph modeling, applying SGT achieves significantly better performance in most cases, regardless of using GPC or not. This demonstrates the effectiveness of the skeleton graph learning with SGT, as it can fully capture relations within skeletons to learn unique body structure and motion features for person re-ID. The SGT employing GPC achieves superior results than "SGT + DS" that uses direct supervised learning (*i.e.*, cross-entropy loss) with an improvement of 4.4-14.9% for mAP and 5.5-24.2% for Rank-1 accuracy, which verifies the key role of the graph prototype contrastive learning in capturing more representative discriminative graph features from different identities. Finally, adding STPR consistently boosts model performance by 1.4-3.7% for mAP on all dataset, which further demonstrates its effectiveness on capturing more valuable graph semantics and discriminative patterns for person re-ID.

5. Further Analysis

Evaluation on RGB-estimated Skeletons. To verify the generality of our skeleton-based model on RGB-estimated skeletons, we utilize pre-trained pose estimation models to extract skeleton data from RGB videos of CASIA-B [8], and evaluate the performance of TranSG under different

Table 4. Performance of our approach on different-scale skeleton graphs. J, 10 or 5 nodes correspond to the joint-scale, part-scale or body-scale skeletal structures used in SM-SGE [19].

Nodes	Methods	BIWI-S		BIWI-W		KS20		KGBD	
		mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
J	SM-SGE [19]	10.0	33.0	14.9	12.9	10.2	44.7	4.3	40.2
	TranSG (Ours)	**30.1**	**68.7**	**26.9**	**32.7**	**46.2**	**73.6**	**20.2**	**59.0**
10	SM-SGE	11.1	32.8	16.5	14.5	9.8	43.2	4.1	33.0
	TranSG (Ours)	**15.1**	**37.9**	**18.7**	**20.0**	**17.0**	**48.1**	**4.9**	**36.9**
5	SM-SGE	10.0	27.5	13.8	12.6	9.3	37.3	4.4	31.5
	TranSG (Ours)	**13.5**	**36.6**	**14.1**	**17.0**	**13.2**	**40.8**	**4.6**	**34.6**

Table 5. Performance of our approach using only unlabeled skeletons. We use the same clustering setting in [11] for comparison.

Methods	BIWI-S		BIWI-W		KS20		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
SPC-MGR [11]	**16.0**	34.1	19.4	18.9	21.7	59.0	6.9	40.8
TranSG (Ours)	14.2	**42.2**	**21.9**	**26.6**	**23.6**	**65.0**	**7.5**	**43.6**

conditions. As presented in Table 3, our approach not only outperforms many existing state-of-the-art skeleton-based models with a prominent improvement in most conditions, but also achieves superior performance to representative classical appearance-based methods that utilize RGB-based features (*e.g.*, textures, silhouettes) or/and visual metric learning [46, 48–51]. This demonstrates the stronger ability of TranSG on capturing more discriminative features from the estimated skeletons, and also suggests its great potential to be applied to more general RGB-based scenarios.

Application to Skeleton Graphs with Varying Scales. To validate the effectiveness of our approach under different graph modeling, we follow [19] to construct different-scale graphs for model learning. As shown in Table 4, TranSG significantly outperforms the state-of-the-art framework SM-SGE [19] when utilizing the original skeleton graphs (corresponding to J joints) or higher level graphs with less nodes. This demonstrates that our model is compatible with different-level skeletal structures and can learn more effective features even under different graph modeling.

Transfer to Unsupervised Scenarios. To apply our approach in an unsupervised manner without using ground-truth labels, we follow [11] to perform DBSCAN clustering [52] of graph representations, and leverage their pseudo classes to generate graph prototypes for contrastive learning. With only *unlabeled* skeletons as inputs, TranSG can still achieve superior performance than the latest graph-based method SPC-MGR [11] in most cases, as shown in Table 5. This further demonstrates the generality of our approach, which could be promisingly transferred to more related tasks such as unsupervised open-set person re-ID.

Discussions. As presented in Fig. 3, we show effects of different hyper-parameters on our approach. An appropriate fusion of different components can encourage better model performance, while setting α, β, and λ to 0.5 achieves slightly better results. Adding too many FR heads or SGT layers could slightly reduce the performance, as it might ex-
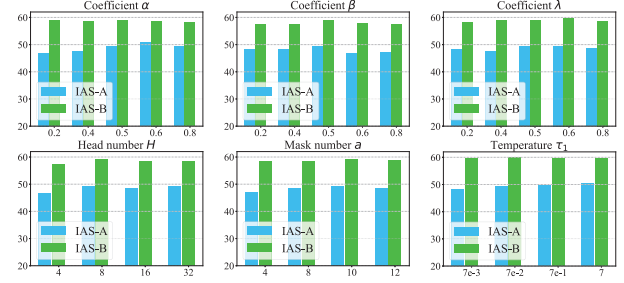

Figure 3. Rank-1 accuracy of our approach on different probe sets (IAS-A and IAS-B) when setting different hyper-parameters.

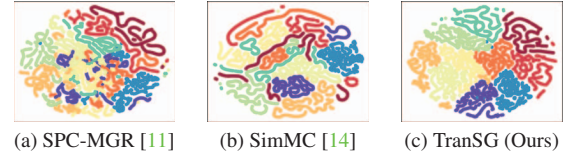
(a) SPC-MGR [11]   (b) SimMC [14]   (c) TranSG (Ours)

Figure 4. t-SNE visualization of representations learned by SPC-MGR (a), SimMC (b), and TranSG (c) for first ten classes in IAS. Different colors indicates representations of different classes.

pand the model scale and learn more redundant information. Our model is not sensitive to the variation of some parameters such as temperature τ_1, while setting a moderate value for mask numbers benefits model performance. The t-SNE visualization [53] in Fig. 4 shows that our learned skeleton representations possess more discriminative inter-class separation than other methods [11, 14], which indicates that TranSG may capture richer class-related semantics. More empirical and theoretical analyses are in the appendices.

6. Conclusion

In this paper, we propose TranSG to learn effective representations from skeleton graphs for person re-ID. We devise a skeleton graph transformer (SGT) to perform full-relation learning of body-joint nodes to aggregate key body and motion features into graph representations. A graph prototype contrastive learning (GPC) approach is proposed to learn discriminative graph representations by contrasting their inherent similarity with the most representative graph features. Furthermore, we design a graph structure-trajectory prompted reconstruction (STPR) mechanism to encourage learning richer graph semantics and key patterns for person re-ID. TranSG outperforms existing state-of-the-art models, and can be scalable to be applied to different scenarios.

Acknowledgements

References

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021. 1

[2] H. Rao, S. Wang, X. Hu, M. Tan, Y. Guo, J. Cheng, X. Liu, and B. 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*, no. 01, pp. 1–1, 2021. 1, 2, 5, 6, 7

[3] A. Nambiar, A. Bernardino, and J. C. Nascimento, "Gait-based person re-identification: A survey," *ACM Computing Surveys*, vol. 52, no. 2, p. 33, 2019. 1

[4] R. Vezzani, D. Baltieri, and R. Cucchiara, "People reidentification in surveillance and forensics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, p. 29, 2013. 1

[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. J. Finocchio, R. Moore, A. A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304, 2011. 1

[6] H. Rao, S. Wang, X. Hu, M. Tan, H. Da, J. Cheng, and B. Hu, "Self-supervised gait encoding with locality-aware attention for person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 1, pp. 898–905, 2020. 1, 2, 6, 7

[7] P. Pala, L. Seidenari, S. Berretti, and A. Del Bimbo, "Enhanced skeleton and face 3D data for person re-identification from depth cameras," *Computers & Graphics*, vol. 79, pp. 69–80, 2019. 1, 2, 6, 7

[8] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognition*, vol. 98, p. 107069, 2020. 1, 2, 6, 7

[9] V. O. Andersson and R. M. Araujo, "Person identification using anthropometric and gait data from Kinect sensor," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 425–431, 2015. 1, 6

[10] M. Munaro, A. Basso, A. Fossati, L. Van Gool, and E. Menegatti, "3D reconstruction of freely moving persons for re-identification with a depth sensor," in *International Conference on Robotics and Automation (ICRA)*, pp. 4512–4519, IEEE, 2014. 1

[11] H. Rao and C. Miao, "Skeleton prototype contrastive learning with multi-level graph relation modeling for unsupervised person re-identification," *arXiv preprint arXiv:2208.11814*, 2022. 1, 2, 5, 6, 7, 8

[12] M. Munaro, A. Fossati, A. Basso, E. Menegatti, and L. Van Gool, "One-shot person re-identification with a consumer depth camera," in *Person Re-Identification*, pp. 161–181, Springer, 2014. 1, 2, 6, 7

[13] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D sensors," in *European Conference on Computer Vision (ECCV) Workshop*, pp. 433–442, Springer, 2012. 1, 2

[14] H. Rao and C. Miao, "SimMC: Simple masked contrastive learning of skeleton representations for unsupervised person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1290–1297, 2022. 1, 2, 5, 6, 7, 8

[15] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017. 1

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. 1

[17] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1229–1238, 2016. 1

[18] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Multi-level graph encoding with structural-collaborative relation learning for skeleton-based person re-identification," in *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 973–980, 2021. 1, 2, 6, 7

[19] H. Rao, X. Hu, J. Cheng, and B. 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*, pp. 1812–1820, 2021. 2, 3, 4, 6, 7, 8

[20] H. Rao, Y. Li, and C. Miao, "Revisiting k-reciprocal distance re-ranking for skeleton-based person re-identification," *IEEE Signal Processing Letters*, vol. 29, pp. 2103–2107, 2022. 2

[21] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3733–3742, 2018. 2

[22] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *International Conference on Learning Representation (ICLR)*, 2021. 2

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020. 2

[24] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020. 2

[25] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021. 2

[26] H. Rao, S. Xu, X. Hu, J. Cheng, and B. Hu, "Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition," *Information Sciences*, vol. 569, pp. 90–109, 2021. 2

[27] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018. 2

[28] F. Petroni, T. Rocktäschel, S. Riedel, P. Lewis, A. Bakhtin, Y. Wu, and A. Miller, "Language models as knowledge bases?," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2463–2473, 2019. 3

[29] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning (ICML)*, pp. 4904–4916, PMLR, 2021. 3

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, pp. 8748–8763, PMLR, 2021. 3

[31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022. 3

[32] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020. 3

[33] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3045–3059, 2021. 3

[34] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4582–4597, 2021. 3

[35] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*, 2021. 3

[36] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4222–4235, 2020. 3

[37] M. P. Murray, A. B. Drought, and R. C. Kory, "Walking patterns of normal men," *Journal of Bone and Joint Surgery*, vol. 46, no. 2, pp. 335–360, 1964. 4

[38] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," in *AAAI Conference on Artificial Intelligence (AAAI) Workshop*, 2021. 4

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017. 4

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. 4

[41] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning (ICML)*, pp. 448–456, PMLR, 2015. 4

[42] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3D skeleton based human motion prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 214–223, 2020. 6

[43] M. Munaro, S. Ghidoni, D. T. Dizmen, and E. Menegatti, "A feature-based approach to people re-identification using skeleton keypoints," in *International Conference on Robotics and Automation (ICRA)*, pp. 5644–5651, IEEE, 2014. 6

[44] A. Nambiar, A. Bernardino, J. C. Nascimento, and A. Fred, "Context-aware person re-identification in the wild via fusion of gait and anthropometric features," in *International Conference on Automatic Face & Gesture Recognition*, pp. 973–980, IEEE, 2017. 6

[45] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *International Conference on Pattern Recognition (ICPR)*, vol. 4, pp. 441–444, IEEE, 2006. 6

[46] Z. Liu, Z. Zhang, Q. Wu, and Y. Wang, "Enhancing person re-identification by integrating gait biometric," *Neurocomputing*, vol. 168, pp. 1144–1156, 2015. 6, 7, 8

[47] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015. 7

[48] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of Machine Learning Research*, vol. 10, no. 2, pp. 207–244, 2009. 7, 8

[49] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *International Conference on Machine Learning (ICML)*, pp. 209–216, 2007. 7, 8

[50] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 262–275, Springer, 2008. 7, 8

[51] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2360–2367, IEEE, 2010. 7, 8

[52] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial

databases with noise.," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, vol. 96, pp. 226–231, 1996. 8

[53] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE.," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008. 8