Motif Guided Graph Transformer with Combinatorial Skeleton Prototype Learning for Skeleton-Based Person Re-Identification

Haocong Rao, Chunyan Miao*

College of Computing and Data Science, Nanyang Technological University (NTU), Singapore
Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), NTU, Singapore
{haocong001,ascymiao}@ntu.edu.sg

Abstract

Person re-identification (re-ID) via 3D skeleton data is a challenging task with significant value in many scenarios. Existing skeleton-based methods typically assume virtual motion relations between all joints, and adopt average joint or sequence representations for learning. However, they rarely explore key body structure and motion such as gait to focus on more important body joints or limbs, while lacking the ability to fully mine valuable spatial-temporal sub-patterns of skeletons to enhance model learning. This paper presents a generic Motif guided graph transformer with Combinatorial skeleton prototype learning (MoCos) that exploits *structure-specific* and *gait-related* body relations as well as combinatorial features of skeleton graphs to learn effective skeleton representations for person re-ID. In particular, motivated by the locality within joints' structure and the body-component collaboration in gait, we first propose the *motif guided graph transformer (MGT)* that incorporates hierarchical structural motifs and gait collaborative motifs, which simultaneously focuses on multi-order local joint correlations and key cooperative body parts to enhance skeleton relation learning. Then, we devise the *combinatorial skeleton prototype learning (CSP)* that leverages random spatial-temporal combinations of joint nodes and skeleton graphs to generate diverse *sub-skeleton* and *sub-tracklet* representations, which are contrasted with the most representative features (*prototypes*) of each identity to learn class-related semantics and discriminative skeleton representations. Extensive experiments validate the superior performance of MoCos over existing state-of-the-art models. We further show its generality under RGB-estimated skeletons, different graph modeling, and unsupervised scenarios.

Code — https://github.com/Kali-Hac/MoCos

Introduction

Person re-identification (re-ID) aims at matching and retrieving a person-of-interest from different views or scenes, which assumes an essential role in security authentication, smart surveillance, human tracking, and robotics (Vezzani, Baltieri, and Cucchiara 2013; Ye et al. 2021). Recent advancements in low-cost and accurate skeleton-tracking devices (*e.g.*, Kinect (Shotton et al. 2011)) have streamlined the collecting of 3D skeletons to make them as a popular and generic data modality for gait analysis and person re-ID (Liao et al. 2020; Rao et al. 2021b; Rao, Leung, and Miao 2024). Compared with traditional person re-ID methods that rely on appearance features (Wang et al. 2016), skeleton-based models typically utilize body structural features and motion patterns of key body joints to identify different persons, which could possess many merits such as lighter inputs and models, better privacy protection (*e.g.*, without using appearances or faces), and more robust performance under view, scale, and background variations (Han et al. 2017).

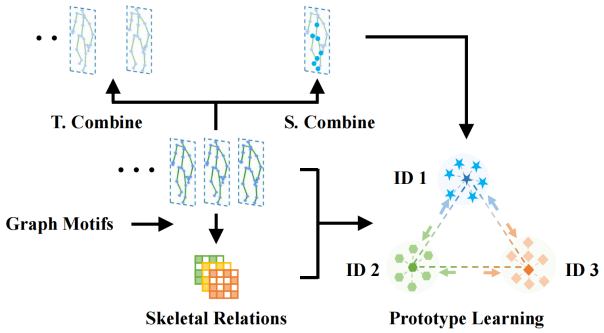


Figure 1: Our approach exploits various graph motifs to enhance skeletal relation learning, and utilizes diverse spatial (S.) and temporal (T.) combinatorial skeleton features to perform skeleton prototype learning for person re-ID.

Early skeleton-based methods (Andersson and Araujo 2015) extract hand-crafted anthropometric and gait descriptors (*e.g.*, kinematic parameters) based on domain expertise, while they are often incapable of exploiting latent skeleton features beyond human cognition. Recent years have witnessed the great success of deep neural networks such as graph transformers for skeleton-based person re-ID (Liao et al. 2020; Rao and Miao 2023). A common practice in these studies is to combine body-joint relation modeling and skeleton prototype learning (*e.g.*, class feature clustering and contrasting) (Rao and Miao 2022; Rao and Miao 2023). However, most methods learn body joint or component relations with the assumption of virtual motion connections among *all* joints (Rao et al. 2021c; Rao and Miao 2023), while they typically lack a specific *focus* on key body

*Corresponding author

joints or local body parts that are highly related to walking patterns ($e.g.$, gait) to capture more discriminative features. On the other hand, existing works usually leverage *average* skeleton or sequential features (Rao and Miao 2022) to perform representation learning, while they rarely harness different *combinatorial* spatial or temporal patterns ($e.g.$, sub-patterns) of key body joints, parts or skeletons to enhance the skeletal structure and motion learning. For example, different combinations of key joints such as hip and knee joints may characterize different structural features in walking patterns, while a combination of partial consecutive skeletons could contain key sub-patterns of a sequence, both of which can be utilized to mine more valuable skeleton features.

To address the aforementioned challenges, we propose a generic **Mo**tif guided graph transformer with **Co**mbinatorial **s**keleton prototype learning (MoCos) (illustrated in Fig. 1), which exploits different *graph motifs* to guide body-joint relation learning in terms of key body structure and motion, and leverages different spatial-temporal feature combinations of both joints and skeletons to enhance skeleton graph representation learning for person re-ID. In particular, motivated by the local correlations (referred to as *locality*) within hierarchical body joints' structure, we first devise ***hierarchical structural motifs (HSM)***, which endow body joints with different semantic roles of connections, to specially focus on multi-order dependencies of body joints and their structural correlations to capture richer skeleton patterns. Then, considering that collaborative movements of upper and lower limbs usually contain unique ($e.g.$, identity-specific) gait patterns (Murray, Drought, and Kory 1964), we propose ***gait collaborative motifs (GCM)*** that focus on both *local* and *global* motion relations of key limbs' joints to encourage the model to capture more salient gait features. By incorporating HSM and GCM into the joint relation learning, we devise the ***motif guided graph transformer (MGT)*** to simultaneously capture key body relations from hierarchical local structure of joints and gait-related collaborative components for person re-ID. Last, to exploit more valuable combinatorial patterns from skeletons and their sequences, a ***combinatorial skeleton prototype learning (CSP)*** approach is proposed to randomly mask body-joint nodes and skeleton graphs to generate spatial-temporal combinatorial graph features at both levels of *sub-skeletons* and *sub-tracklets*, which are utilized to contrast and learn the most representative skeleton graph features (referred to as *prototypes*) of each identity. CSP pulls different combinatorial skeleton graph representations closer to corresponding prototypes, and pushes them apart from other prototypes, so as to facilitate the model to learn distinguishing skeleton features and high-level class-related semantics for person re-ID.

Our main contributions can be summarized as follows:

- We propose a generic MoCos paradigm that exploits diverse graph motifs and combinatorial skeleton features to learn effective representations from skeleton graphs for person re-ID. To the best of our knowledge, MoCos is the first exploration of structure-specific and gait-based graph motifs to enhance skeleton relation and prototype learning specifically for skeleton-based person re-ID.
- We devise the motif guided graph transformer (MGT) by synergizing hierarchical structural motifs (HSM) and gait collaborative motifs (GCM) to guide body-joint relation learning, so as to capture more discriminative body structural and gait features within skeletons for person re-ID.
- We propose the combinatorial skeleton prototype learning (CSP) that leverages combinatorial spatial-temporal graph features of joints (sub-skeletons) and skeletons (sub-tracklets) to learn more key skeleton patterns.
- Empirical evaluations on five public datasets validate that MoCos significantly outperforms existing state-of-the-art methods and can be effectively applied to different graph modeling, RGB-estimated or unsupervised scenarios.

Related Works

Skeleton-Based Person Re-Identification. Skeleton-based person re-ID focuses on the problem of matching and retrieving a certain person based on spatial and temporal representations of skeletal human body and gait (2024; 2024; 2023; 2024). Early-stage studies manually extract skeleton or body-joint descriptors in terms of anthropometric and gait attributes for person re-ID. Barbosa et al. compute Euclidean distances between different joint pairs as descriptors, while they are further extended to 13 (D_{13} (Munaro et al. 2014b)) and 16 skeleton descriptors (D_{16} (Pala et al. 2019)) to perform person re-ID. Most recent methods (2020; 2021b; 2023; 2024) leverage deep learning models for skeleton sequence or skeleton graph representation learning. PoseGait (Liao et al. 2020) is proposed to encode 3D pose features and joint-based motion descriptors (denoted as D_{PG}) for human recognition. Rao et al. utilize an encoder-decoder model with attention mechanisms (AGE) to encode skeleton-based gait patterns, while its extension SGELA (Rao et al. 2021b) further enhances self-supervised skeleton semantic learning with diverse skeletal pretext tasks (e.g., time series forecasting (Feng et al. 2024; Zhicheng et al. 2024)) and inter-sequence contrastive mechanisms for the person re-ID task. Rao and Miao propose a masked contrastive learning framework (SimMC) to perform skeleton prototype learning with intra-sequence relation learning for person re-ID. The multi-scale skeleton graphs are explored in (2021c; 2021a; 2022) to learn body relations and patterns at various levels. In (Rao and Miao 2023), a skeleton graph transformer is devised to learn both skeleton and sequential graph features for person re-ID. A general skeleton feature re-ranking mechanism is proposed in (Rao, Li, and Miao 2022) for skeleton-based person re-ID. Hi-MPC (Rao, Leung, and Miao 2024) utilizes hierarchical prototype learning with a hard skeleton mining approach to learn discriminative skeleton features. Existing multi-modal person re-ID methods usually combine skeleton-based features with extra RGB or depth information ($e.g.$, depth shape features based on point clouds (Munaro et al. 2014a; Hasan and Babaguchi 2016; Wu, Zheng, and Lai 2017)) to boost re-ID accuracy. For example, some works combine RGB images and skeleton data to learn auxiliary anthropometric attributes (Wang et al. 2020), body parts correlations (Lu et al. 2023), and clothing-invariant features (Nguyen et al. 2024) to enhance their performance.

Graph Motifs. Motifs define different patterns of connections in graphs or networks via specifying the pattern-context nodes relevant to a target node of interest (Sankar, Zhang, and Chang 2017), which have been widely applied to many areas such as neuroscience and computer vision. (2004; 2007; 2022; 2023) A few recent works (Wen et al. 2022, 2019) integrate motifs into graph convolutional networks (GCNs) to learn skeleton features from joints of interest and their context for action recognition. As far as we know, this work is *the first exploration* of structural and gait-based graph motifs with high-order semantic roles of joints *specifically* for skeletal relation learning and person re-ID.

Methodology

Preliminary

Problem Definition. Suppose that a 3D skeleton sequence $\boldsymbol{X}=(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_f)\in\mathbb{R}^{f\times J\times 3}$, where f denotes the number of skeletons in the sequence and $\boldsymbol{x}_i\in\mathbb{R}^{J\times 3}$ represents the i^{th} skeleton with 3D coordinates of J body joints. Each sequence $\boldsymbol{X}$ corresponds to an identity class y $\in\{1,\cdots,C\}$ and C is the number of different identity classes. We denote the Training set, Probe set, and Gallery set as $\Phi_{\mathcal{T}}=\left\{\boldsymbol{X}_i^{\mathcal{T}}\right\}_{i=1}^{n_1}$, $\Phi_{\mathcal{P}}=\left\{\boldsymbol{X}_i^{\mathcal{P}}\right\}_{i=1}^{n_2}$, and $\Phi_{\mathcal{G}}=\left\{\boldsymbol{X}_i^{\mathcal{G}}\right\}_{i=1}^{n_3}$, which respectively contain n_1, n_2, and n_3 skeleton sequences of different persons collected from different scenes or views. The model target is to encode skeleton sequences into effective representations, so that we can query the correct identity of each skeleton sequence representation (denoted as $\{\boldsymbol{V}_i^{\mathcal{P}}\}_{i=1}^{n_2}$) in the probe set via matching it with the sequence representations (denoted as $\{\boldsymbol{V}_i^{\mathcal{G}}\}_{i=1}^{n_3}$) in the gallery set.

Skeleton Graph Construction. We construct skeleton graphs based on the physical connections of human body joints (Rao and Miao 2023): For the t^{th} skeleton $\boldsymbol{x}_t$, we represent it as the graph $\mathcal{G}^t(\mathcal{V}^t,\mathcal{E}^t)$, which consists of J 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 $\mathcal{E}^t=\{e_{i,j}^t\,|\boldsymbol{v}_i^t,\boldsymbol{v}_j^t\in\mathcal{V}^t\}$, $e_{i,j}^t\in\mathbb{R}$. Here $\mathcal{E}^t$ denotes the set of connections and motion relations between different joints, and can be represented with an adjacent matrix $\mathbf{A}^t\in\mathbb{R}^{J\times J}$, initialized by the connections of adjacent body joints.

Motif Guided Graph Transformer

Different body joints and parts of a pedestrian typically possess unique relations, such as structural relations between adjacent joints, and actional relations between non-adjacent parts, characterizing discriminative walking patterns (Murray, Drought, and Kory 1964; Rao et al. 2021a). Existing methods typically perform global relation learning with the assumption of virtual motion relations among all joints (Rao and Miao 2023), while they rarely exploit local hierarchical structure of joints (defined as "*locality*") or key gait-related body components to capture richer valuable relations. To this end, we propose to endow body-joint nodes with different relational semantic roles (defined as "*motifs*" for skeleton graphs), and devise the ***motif guided graph transformer (MGT)*** to simultaneously focus on their *hierarchical structural relations* and *gait collaborative relations* to learn effective skeleton graph representations for person re-ID.

Graph Transformer (GT). First, given a skeleton graph, its J node representations are integrated with their positional encoding based on the graph adjacency matrix $\mathbf{A}^t$ (Rao and Miao 2023), which can be formulated as:

$$\boldsymbol{h}_i=(\mathbf{W}_1\boldsymbol{v}_i+\boldsymbol{b}_1)+(\mathbf{W}_2\boldsymbol{\lambda}_i+\boldsymbol{b}_2)\,, \tag{1}$$

where $\boldsymbol{h}_i\in\mathbb{R}^D$ represents the position-encoded representation of i^{th} node, $\lambda_i\in\mathbb{R}^K$ denotes the i-node's positional encoding extracted from the K smallest non-trivial eigenvectors of graph Laplacian matrix following (Dwivedi and Bresson 2021), and $\mathbf{W}_1\in\mathbb{R}^{D\times 3},\mathbf{W}_2\in\mathbb{R}^{D\times K},\boldsymbol{b}_1,\boldsymbol{b}_2\in\mathbb{R}^D$ are learnable parameters to map i^{th} node $\boldsymbol{v}_i$ and corresponding positional encoding into feature spaces of the same dimension D. Then, GT computes the *preliminary* relation value of joints (referred to as "*full relations (FR)*") by

$$\boldsymbol{R}_{i,j}^{k,l}=\mathrm{Softmax}_j\left(\frac{(\boldsymbol{Q}^{k,l}\boldsymbol{h}_i^{(l)})\cdot(\boldsymbol{K}^{k,l}\boldsymbol{h}_j^{(l)})}{\sqrt{D_{\mathrm{k}}}}\right). \tag{2}$$

In Eq. (2), $\boldsymbol{Q}^{k,l},\boldsymbol{K}^{k,l}\in\mathbb{R}^{D_{\mathrm{k}}\times D}$ represent the learnable weight matrices for query and key transformations in the k^{th} relation head of the l^{th} GT layer, $\frac{1}{\sqrt{D_{\mathrm{k}}}}$ is the scaling factor of dot-product similarity, and $\boldsymbol{R}_{i,j}^{k,l}$ denotes the $\mathrm{softmax}$-normalized relational value between the i^{th} and j^{th} joint captured by the k^{th} relation head in the l^{th} layer.

Hierarchical Structural Motifs. To guide the model to fully capture skeleton patterns from body joints' physical connections and the multi-level dependencies within their local hierarchical structure, we devise the *hierarchical structural motifs (HSM)* to learn structural body relations from different-order neighbors of joint nodes. The focused body-joint relations of HSM can be represented as a matrix with

$$\mathcal{A}_{i,j}^m=\begin{cases}1 & \text{if } j\in\bigcup_{k=1}^{m}\mathcal{N}_i^k\\ 0 & \text{otherwise}\end{cases}, \tag{3}$$

where $m\in\{1,2,3\}$, $\mathcal{A}^m\in\mathbb{R}^{J\times J}$ denote the m-order HSM matrix, $i\in\{1,2,\cdots,J\}$, and $\mathcal{N}_i^k$ represents the indices for k-order neighbors of the i^{th} body-joint node ($i.e.$, nodes with k-hop distance to the i^{th} node). Intuitively, the m-order HSM $\mathcal{A}^m$ defines $R_m=2m+1$ semantic roles for all joint nodes: $\mathcal{A}^1$ contains $R_1=3$ roles, including a joint node itself, its parent node, and child node; $\mathcal{A}^2$ contains $R_2=5$ roles, including a joint node itself, its grandparent node, parent node, child node, and grandchild node, while $\mathcal{A}^3$ ($R_3=7$) further includes the roles of its great-grandparent node and great-grandchild node. Note that HSM does NOT require pre-defining the directions of node connections but views them as bi-directional to focus on the general hierarchical structure of joints. The maximum order is empirically set to 3 as the center joint of spine in most datasets has up to 3-hop neighbors (Li et al. 2021). By simultaneously focusing on relations of each body-joint node to its immediate and higher-level connected neighbors, HSM encourages the model to encode the inherent hierarchical structure ($e.g.$, high-order locality) of joints' positions and motion to capture more valuable patterns of skeleton graphs.

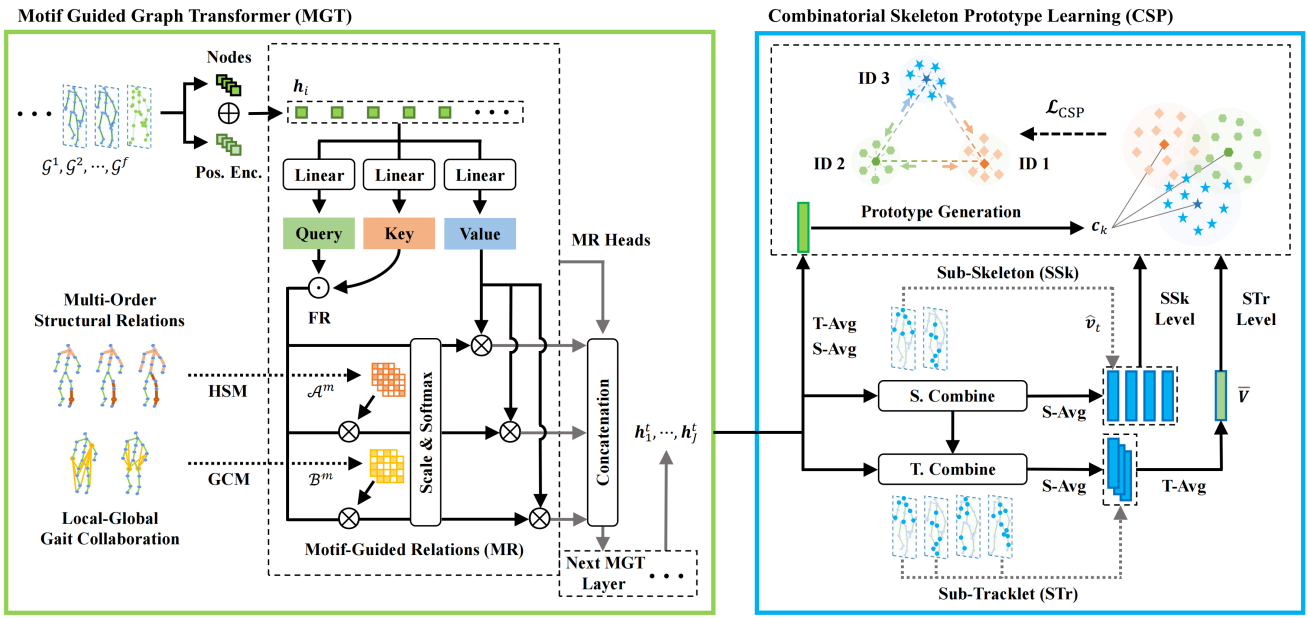


Figure 2: Schematics of our approach: First, with position-encoded node representations for each skeleton graph $\mathcal{G}^t$, MGT incorporates hierarchical structural motifs (HSM) and gait collaborative motifs (GCM) to perform body relation learning, which concurrently focuses on multi-order structural correlations and gait-related collaborative body parts to enhance skeleton pattern learning. Then, CSP temporally and spatially masks joints and graphs to generate combinatorial sub-skeleton (SSk) and sub-tracklet (STr) representations, which are contrasted with skeleton prototypes generated from same-identity spatially-temporally averaged (S-Avg and T-Avg) skeleton graph representations. We enhance the similarity of both SSk and STr level features to their corresponding prototypes, while maximizing their dissimilarity to other prototypes by optimizing $\mathcal{L}_{\text{CSP}}$.

Gait Collaborative Motifs. Motivated by the gait property that different key body components ($e.g.$, arms and legs) usually perform collaborative motion characterizing identity-specific patterns (Murray, Drought, and Kory 1964), we propose the *gait collaborative motifs (GCM)* to guide the model to learn more salient patterns from motion units of both upper and lower limbs (see Fig. 2). In particular, we regard each body joint in limbs as a basic motion unit, and focus on its local relations *within the same limb* and global relations *with other limbs* to facilitate the gait pattern learning. We define GCM with the focused body-joint relations as

$$\mathcal{B}^m_{i,j} = \begin{cases} 1 & \text{if } i \in \mathcal{I}^m,\ j \in \bigcup_{k=1}^{2} \mathcal{I}^k,\ j \neq i \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $m \in \{1, 2\}$, $\mathcal{B}^1, \mathcal{B}^2 \in \mathbb{R}^{J \times J}$ denote the GCM matrices for the upper and lower limbs, $\mathcal{I}^1$ and $\mathcal{I}^2$ represent the sets of indices for joint nodes in upper limbs ($e.g.$, arms) and lower limbs ($e.g.$, legs) respectively (visualized in Appendix I). Specifically, each GCM matrix defines $\hat{R} = 3$ semantic roles for all joint nodes: $\mathcal{B}^1$ (or $\mathcal{B}^2$) contains the roles of a joint node in a upper (or lower) limb, its *locally*-correlated sibling nodes ($i.e.$, nodes in the same limb), and its *globally*-collaborative nodes in other limbs. In this way, GCM aims to focus on both local and global relation learning of limb motion units to encourage mining more unique cooperative skeleton patterns from their gait-related components.

By incorporating HSM and GCM into the relation learning process of GT, we devise the *motif guided graph transformer (MGT)* to *jointly* focus on hierarchical body-joint structure and gait-related components to capture more key skeleton patterns. In particular, MGT computes *motif-guided relations (MR)* by updating Eq. (2) to (illustrated in Fig. 2)

$$\hat{\boldsymbol{R}}^{k,l}_{i,j} = \text{Softmax}_j\left(\frac{\mathcal{M}^k_{i,j}(\boldsymbol{Q}^{k,l}\boldsymbol{h}^{(l)}_i)\cdot(\boldsymbol{K}^{k,l}\boldsymbol{h}^{(l)}_j)}{\sqrt{D_{\text{k}}}}\right), \tag{5}$$

where

$$\mathcal{M}^k_{i,j} = \begin{cases} \mathcal{A}^k_{i,j} & \text{if } k \in \{1,2,3\} \\ \mathcal{B}^{k-3}_{i,j} & \text{if } k \in \{4,5\} \\ 1 & \text{otherwise} \end{cases}. \tag{6}$$

In Eq. (5) and (6), $\boldsymbol{h}^{(l)}_i \in \mathbb{R}^D$ denotes the feature representation of the i^{th} joint encoded by the l^{th} MGT layer, $\hat{\boldsymbol{R}}^{k,l}_{i,j}$ represents the relation value between the i^{th} and j^{th} joint computed by the k^{th} MR head in the l^{th} layer, $k \in \{1, 2, \cdots, H\}$, and H is the number of MR heads. MGT adopts multiple MR heads to jointly perform motif-guided and full relation learing, which are then aggregated into the each graph node representation with

$$\hat{\boldsymbol{h}}^{(l)}_i = \boldsymbol{O}^l \Big\|_{k=1}^{H} \left(\sum_{j=1}^{J} \hat{\boldsymbol{R}}^{k,l}_{i,j}\boldsymbol{V}^{k,l}\boldsymbol{h}^{(l)}_j\right), \tag{7}$$

where $\boldsymbol{V}^{k,l} \in \mathbb{R}^{D_k \times D}$ represent the learnable weight matrices for value transformation in the k^{th} MR head of the l^{th} MGT layer, $\boldsymbol{O}^l \in \mathbb{R}^{D \times D}$ is the parameter matrix for output transformation, $\Big\|$ represents the concatenation operation, $\hat{\boldsymbol{h}}_i^{(l)} \in \mathbb{R}^D$ denotes the i^{th} node representation that concatenates node features learned from different MR heads in the l^{th} layer, For convenience, we use $\boldsymbol{h}_i^t$ to denote the final representation ($i.e.$, concatenated node representation of the last MGT layer) of i^{th} node in t^{th} skeleton graph.

By integrating different motifs (Eq. (3), (4)) into the relation computation (Eq. (5), (6)), we encourage the model to focus on both multi-level structural relations and gait-related collaboration of key joints to capture richer effective patterns for person re-ID. It is worth noting that the proposed multi-head MGT naturally generalizes the self-attention based GT (Rao and Miao 2023) to local and global body relation learning using skeleton-specific motifs. The proposed motifs can also be generally applied to non-graph models, unsupervised skeleton data, and different-scale skeleton representations.

Combinatorial Skeleton Prototype Learning

To mine the most representative skeleton features of each identity for person re-ID, existing solutions (Rao and Miao 2022; Rao and Miao 2023) typically *average* spatial or temporal features of skeletons for prototype clustering and contrasting, while they rarely harness different *combinatorial* spatial-temporal patterns of body joints, parts or skeletons to learn more effective representations. For example, a subset or dynamic combination of key body joints such as wrist, knee and foot joints (defined as "*sub-skeleton representations*") can depict different body structural features within gait, while different key segments of a skeletal walking tracklet (defined as "*sub-tracklet representations*") typically contain diverse sub-patterns (Zhang et al. 2020), both of which can be exploited to learn more informative and unique features. To this end, we propose the ***combinatorial skeleton prototype learning (CSP)*** that leverages spatial-temporal combinatorial graph representations of sub-skeletons and sub-tracklets to jointly perform skeleton prototype learning.

Given the t^{th} skeleton graph representation $(\boldsymbol{h}_1^t, \cdots, \boldsymbol{h}_J^t)$ containing J spatial representations of body-joint nodes, we utilize random masks to generate a subset of nodes to construct its spatial combinatorial representation ($i.e.$, sub-skeleton representation) by

$$\hat{\boldsymbol{v}}_t = \frac{1}{N_S} \sum_{j=1}^{J} x_j \boldsymbol{h}_j^t, \tag{8}$$

where $\hat{\boldsymbol{v}}_t \in \mathbb{R}^D$ is the sub-skeleton representation of t^{th} skeleton graph by randomly masking nodes, $x_j \in \{0, 1\}$ denotes the j^{th} mask that is an independent and identically distributed Bernoulli random variable with the probability p_s of being 0 ($i.e.$, $x_j \sim \text{Bernoulli}(1 - p_s)$), and $N_S = \sum_{j=1}^{J} x_j$ represents the number ($i.e.$, subset size) of unmasked node representations. Each unmasked node representation is assumed to be equally important and we average them to be the sub-skeleton graph representation. In practice, the maximum number of masked node representations is $J - 1$ ($i.e.$, $N_S \geq 1$) to avoid empty skeleton representation. Here we adopt Bernoulli distribution for combinatorial feature generation due to its simplicity and computational tractability (Boluki et al. 2020), while other probabilistic distributions can be also extended and applied to the proposed masking.

Then, provided the spatial combinatorial representations $(\hat{\boldsymbol{v}}_1, \cdots, \hat{\boldsymbol{v}}_f)$ of f consecutive skeleton frames (defined as "*a skeletal walking tracklet*"), we generate a random subset of the walking tracklet to yield the spatial-temporal combinatorial representation ($i.e.$, sub-tracklet representation) with

$$\overline{\boldsymbol{V}} = \frac{1}{N_T} \sum_{t=1}^{f} m_t \hat{\boldsymbol{v}}_t, \tag{9}$$

where $\overline{\boldsymbol{V}} \in \mathbb{R}^D$ denotes the sub-tracklet representation that incorporates both spatial and temporal combinatorial features of a skeleton sequence, and $m_t \in \{0, 1\}$ represents the t^{th} random mask sampled from the Bernoulli distribution with the probability p_t being 0. $N_T = \sum_{t=1}^{f} m_t$, $N_T \geq 1$ is the sequence length of sub-tracklet. Each sub-skeleton representation within the sub-tracklet is assigned with the same importance, and we average them as the final sub-tracklet representation. It is worth noting that a sub-tracklet contains sub-trajectory of partial body joints ($i.e.$, sub-skeletons), and can be regarded as a subset representation of sub-sequence trajectory. In essence, the temporally-masked sequence in SimMC (Rao and Miao 2022) and average spatially-masked skeleton representations in TranSG (Rao and Miao 2023) can be viewed as two special cases of proposed sub-tracklet representation by setting $p_s = 0$ and $p_t = 0$ respectively.

To exploit graph representations of both sub-skeletons and sub-tracklets to learn the most discriminative skeleton graph features (defined as "*prototypes*") of each person and high-level semantics ($e.g.$, identity-associated patterns), we propose the combinatorial skeleton prototype (CSP) loss as

$$\mathcal{L}_{\text{CSP}} = \lambda \mathcal{L}_{\text{CSP}}^{str} + (1 - \lambda) \mathcal{L}_{\text{CSP}}^{ssk}, \tag{10}$$

where

$$\mathcal{L}_{\text{CSP}}^{str} = \frac{1}{n_1} \sum_{i=1}^{n_1} -\log \frac{\exp\left(\overline{\boldsymbol{V}}_i \cdot \overline{\boldsymbol{c}} / \tau_1\right)}{\sum_{k=1}^{C} \exp\left(\overline{\boldsymbol{V}}_i \cdot \boldsymbol{c}_k / \tau_1\right)}, \tag{11}$$

$$\mathcal{L}_{\text{CSP}}^{ssk} = \frac{1}{f n_1} \sum_{i=1}^{n_1} \sum_{t=1}^{f} -\log \frac{\exp\left(\mathcal{F}_1\left(\hat{\boldsymbol{v}}_t^i\right) \cdot \mathcal{F}_2\left(\hat{\boldsymbol{c}}\right) / \tau_2\right)}{\sum_{k=1}^{C} \exp\left(\mathcal{F}_1\left(\hat{\boldsymbol{v}}_t^i\right) \cdot \mathcal{F}_2\left(\boldsymbol{c}_k\right) / \tau_2\right)}, \tag{12}$$

$$\boldsymbol{c}_k = \frac{1}{u_k} \sum_{\text{y}_j = k} \boldsymbol{V}_j. \tag{13}$$

The proposed CSP loss in Eq. (10) combines both *sub-tracklet-level* ($\mathcal{L}_{\text{CSP}}^{str}$) and *sub-skeleton-level* combinatorial prototype loss ($\mathcal{L}_{\text{CSP}}^{ssk}$) with the fusion coefficient λ. In Eq. (11), (12) and (13), n_1 is the number of training skeleton sequences, $\hat{\boldsymbol{v}}_t^i$ and $\overline{\boldsymbol{V}}_i$ denote the sub-skeleton representation of the t^{th} skeleton (see Eq. (8)) and the sub-tracklet representation of the i^{th} skeleton sequence (see Eq. (9)). $\overline{\boldsymbol{c}}$ and $\hat{\boldsymbol{c}}$ correspond to their prototypes ($i.e.$, class feature centroids) generated by averaging all sequence representations of the same identity (see Eq. (13)), $\boldsymbol{c}_k$ is the skeleton prototype of

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}
H.	D_{PG} (2020)	6.7	18.5	45.4	63.8	8.7	6.5	15.5	20.3	11.3	35.2	61.5	70.5	11.0	16.4	39.5	53.4	10.6	16.0	41.2	57.3	2.1	30.0	49.1	58.1
	D_{13} (2014b)	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} (2019)	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
S.	PoseGait (2020)	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 (2020)	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 (2021b)	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
	SimMC (2022)	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
	Hi-MPC (2024)	17.4	47.5	70.3	78.6	22.6	27.3	40.3	48.8	22.0	69.6	83.5	87.1	23.2	45.6	67.3	75.4	25.3	48.2	70.2	77.8	10.2	56.9	70.2	75.1
G.	MG-SCR (2021c)	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 (2021a)	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 (2022)	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
	ST-GCN (2018)	28.5	61.6	78.2	89.5	28.2	32.9	47.6	54.8	40.1	60.4	79.9	84.6	34.0	41.6	60.6	68.2	28.1	49.1	68.1	76.3	21.1	57.7	71.6	77.2
	TranSG (2023)	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
	MoCos (Ours)	**32.1**	**72.0**	**89.5**	**93.0**	**30.5**	**36.0**	**49.2**	**57.0**	**50.8**	**76.0**	**87.3**	**90.2**	**35.8**	**51.9**	**69.4**	**77.5**	**45.5**	**61.5**	**79.1**	**87.8**	**26.1**	**62.0**	**75.2**	**79.6**

Table 1: Person re-ID performance comparison with state-of-the-art **H**and-crafted methods (**H.**), **S**equence representation learning methods (**S.**), and **G**raph-based methods (**G.**). **Bold numbers** denote the best performance results among all methods.

k^{th} class, u_k denotes the number of skeleton sequence representations $\boldsymbol{V}_j$ with the class label $\mathrm{y}_j = k$, and τ_1, τ_2 represent temperatures for contrastive learning. $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ are learnable projections to transform sequence-level prototypes and sub-skeleton-level features into the same feature space and integrate related features for contrastive learning. $\mathcal{L}_{\text{CSP}}$ can be viewed as a generalized skeleton prototype loss that incorporates *joint-level* motif-guided relation learning, *sub-skeleton-level* and *sub-tracklet-level* prototype contrasting to enhance spatial-temporal skeleton pattern learning, which can be theoretically modeled as a generalized Expectation-Maximization (EM) solution (see Appendix II).

Experiments

Experimental Setups

Datasets. Four skeleton-based person re-ID benchmark datasets are used to evaluate our approach, including *IAS* (Munaro et al. 2014c), *KS20* (Nambiar et al. 2017), *BIWI* (Munaro et al. 2014b), *KGBD* (Andersson and Araujo 2015), which contain 11, 20, 50, and 164 different persons. The generality of MoCos is also validated on a large-scale multi-view gait dataset *CASIA-B* (Yu, Tan, and Tan 2006) with RGB-estimated skeleton data of 124 individuals under three conditions (Normal (N), Bags (B), Clothes (C)). The commonly-used standard probe and gallery settings (Rao and Miao 2023) are adopted for a fair comparison.

Implementation Details. The skeletons in KGBD, IAS, and BIWI contain $J = 20$ body joints, while KS20 and CASIA-B (RGB-estimated skeletons) contain $J = 25$ and $J = 14$ joints, respectively. For a fair comparison, we follow existing methods (Rao, Leung, and Miao 2024) to set the sequence length to $f = 6$ for IAS, KS20, BIWI, KGBD and $f = 40$ for the RGB-estimated skeleton data in CASIA-B. We set the embedding size to $D = 128$ for each node representation, and empirically employ 2 MGT layers with $H = 8$ relation heads and $D_{\text{k}} = 16$ for each layer. The probability for spatial or temporal masking of CSP is empirically set for different datasets: $p_s = 0.25$, $p_t = 0.25$ for IAS, BIWI, KS20, and $p_s = 0.5$, $p_t = 0.25$ for KGBD. We use fusion coefficient $\lambda = 0.9$ for BIWI-W, KGBD, KS20, $\lambda = 0.25$ for BIWI-S, $\lambda = 0.75$ for IAS-A and IAS-B. We set the learning rate to 3.5×10^{-4} and use an Adam optimizer with batch size 256 for model training on all datasets. More technical details are provided in the appendices.

Evaluation Metrics. Cumulative matching characteristics curve is computed and we report Rank-1, Rank-5, Rank-10 accuracy (R_1, R_5, R_{10}), and Mean Average Precision (mAP) (Zheng et al. 2015) to evaluate model performance.

Comparison with State-of-the-Art Methods

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

Comparison with Graph-based Methods: As shown in Table 1, the proposed MoCos significantly outperforms existing state-of-the-art graph-based methods (SPC-MGR (Rao and Miao 2022), MG-SCR (Rao et al. 2021c), SM-SGE (Rao et al. 2021a)) with an improvement of 11.1-41.3% for mAP and 10.0-51.9% for Rank-1 accuracy on different benchmark datasets. Unlike these methods that resort to multi-scale graph modeling and multi-stage relation learning, our approach can utilize simpler *single-level* graph representations with motif guided *concurrent* relation learning (MGT) to more effectively capture distinguishing skeleton features for person re-ID. In contrast to the latest skeleton graph model TranSG (Rao and Miao 2023) employing naïve GT, our model using MGT also consistently achieves better performance in terms of mAP (2.0-6.1%), Rank-1 (2.4-3.3%), Rank-5 accuracy (0.9-4.3%), and Rank-10 accuracy (0.0-4.8%) on all datasets. This demonstrates the higher efficacy of MoCos incorporating joint-level motif-guided relation learning and different-level prototypical contrast (CSP) to learn richer unique skeleton features for person re-ID. We will also discuss its generality under diverse graph modeling and different unsupervised paradigms in the next section.

Comparison with Hand-crafted and Sequence Learning Methods: Compared with methods that rely on hand-crafted pose features (D_{PG} (Liao et al. 2020)) or anthropo-

ID	GT	HSM	GCM	CSP	BIWI-S		BIWI-W		KS20		KGBD	
					R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
1					38.1	11.3	21.2	18.3	64.8	20.5	53.0	11.0
2	✓				66.6	26.7	31.2	25.5	71.3	42.5	57.0	18.1
3	✓	✓			69.0	29.1	33.0	27.1	74.5	48.4	59.3	24.1
4	✓		✓		69.4	29.6	34.0	28.2	74.4	48.9	60.2	24.0
5	✓	✓	✓		70.8	31.4	34.5	29.4	75.2	50.1	60.9	25.8
6	✓	✓	✓	✓	72.0	32.1	36.0	30.5	76.0	50.8	62.0	26.1

Table 2: Ablation study on different components: Graph transformer (GT), hierarchical structural motifs (HSM), gait collaborative motifs (GCM), and combinatorial skeleton prototype learning (CSP). ✓indicates using the component.

metric attributes (D_{13} (Munaro et al. 2014b), D_{16} (Pala et al. 2019)), our approach achieves superior performance by a marked margin of up to 53.5% Rank-1 accuracy and 39.5% mAP on different benchmarks. Moreover, MoCos also obtains significantly higher performance than latest skeleton sequence contrastive models SimMC (Rao and Miao 2022) and Hi-MPC (Rao, Leung, and Miao 2024) that utilize temporally-masked or hierarchical skeleton representations. This demonstrates the stronger effectiveness of our skeleton graph contrastive model (CSP) that combines both sub-skeleton and sub-tracklet level spatial-temporal features to capture more recognizable patterns for person re-ID.

Ablation Study

We conduct ablation study to evaluate the effectiveness and contribution of each component in our approach. As shown in Table 2, we adopt the direct prototype learning (DP) of skeleton sequences as the baseline (ID = 1) and include GT with direct graph prototype learning (Rao and Miao 2023) (ID = 2) for comparison. In contrast to DP or GT without employing relation learning or graph motifs, integrating motifs HSM or GCM into the body-joint relation learning obtains significantly higher mAP (1.6-28.4%) and Rank-1 accuracy (1.8-31.7%) on different datasets, while combining them (MGT) (ID = 5) further improves the overall performance. This demonstrates the effectiveness of both HSM and GCM, as they can function individually or be compatibly combined to capture more discriminative relational features from structural and gait aspects for person re-ID. Furthermore, incorporating combinatorial skeleton prototype learning (CSP) into MGT (ID = 6) consistently achieves higher results by up to 1.5% for Rank-1 accuracy and 1.1% for mAP on all datasets. This verifies the ability of CSP to utilize spatially-temporally combined sub-skeleton and sub-tracklet representations to enhance the capture of key skeleton patterns and class-related semantics for person re-ID.

Further Analysis

Application to RGB-estimated Scenarios. To verify the generality of MoCos on RGB-estimated skeletons, we extract skeleton data with pre-trained pose estimation models (Cao et al. 2019; Chen and Ramanan 2017) from RGB videos instead of depth sensors. The results in Table 3 show that our model not only achieves superior performance to most existing state-of-the-art skeleton-based models, but also outperforms many representative established appearance-based methods that rely on RGB-based features ($e.g.$, silhouettes) or/and visual metric learning (Liu et al. 2015; Farenzena et al. 2010). This verifies the generality and higher effectiveness of MoCos to learn discriminative patterns from estimated skeletons, and demonstrates its potential for person re-ID under large-scale RGB-based scenarios.

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
A.	LMNN (2009)	—	3.9	—	18.3	—	17.4	—	11.6	—	23.1
	ITML (2007)	—	7.5	—	19.5	—	20.1	—	10.3	—	21.8
	ELF (2008)	—	12.3	—	5.8	—	19.9	—	5.6	—	17.1
	SDALF (2010)	—	4.9	—	10.2	—	16.7	—	11.6	—	22.9
	MLR (Features)	—	13.6	—	13.6	—	13.5	—	9.7	—	14.7
	MLR (Scores) (2015)	—	16.3	—	18.9	—	25.4	—	20.3	—	31.8
S.	AGE (2020)	3.5	20.8	9.8	37.1	9.6	35.5	3.0	14.6	3.9	32.4
	SM-SGE (2021a)	6.6	50.2	9.3	26.6	9.7	27.2	3.0	10.6	3.5	16.6
	SPC-MGR (2022)	9.1	71.2	11.4	44.3	11.8	48.3	4.3	22.4	4.6	28.9
	SGELA (2021b)	9.8	71.8	16.5	48.1	7.1	51.2	4.7	15.9	6.7	36.4
	SimMC (2022)	10.8	84.8	16.5	69.1	15.7	68.0	5.4	25.6	7.1	42.0
	TranSG (2023)	13.1	78.5	17.9	67.1	15.7	65.6	6.7	23.0	8.6	44.1
	Hi-MPC (2024)	11.2	85.5	17.0	71.2	14.1	70.2	4.9	**27.2**	7.5	50.1
	MoCos (Ours)	**16.1**	**87.9**	**18.9**	**73.6**	**18.1**	**72.1**	**7.3**	26.5	**9.8**	**50.6**

Table 3: Person re-ID performance comparison with **A**ppearance-based (**A.**) or **S**keleton-based (**S.**) methods on CASIA-B. "**C-N**" denotes using "**C**lothes" probe set and "**N**ormal" gallery set. "—" indicates no published result.

Scales	Methods	BIWI-S		BIWI-W		KS20		KGBD	
		R_1	mAP	R_1	mAP	R_1	mAP	R_1	mAP
J-Scale	SM-SGE (2021a)	33.0	10.0	12.9	14.9	44.7	10.2	40.2	4.3
	MoCos (Ours)	**72.0**	**32.1**	**36.0**	**30.5**	**76.0**	**50.8**	**62.0**	**26.1**
P-Scale	SM-SGE	32.8	11.1	14.5	16.5	43.2	9.8	33.0	4.1
	MoCos (Ours)	**38.3**	**14.7**	**20.7**	**19.0**	**49.0**	**15.7**	**35.9**	**4.7**
B-Scale	SM-SGE	27.5	10.0	12.6	13.8	37.3	9.3	**31.5**	**4.4**
	MoCos (Ours)	**35.6**	**12.5**	**18.4**	**16.9**	**41.1**	**13.5**	30.6	**4.4**

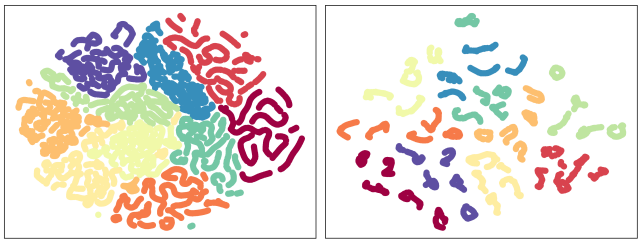
Table 4: Performance of MoCos on **J**oint (**J**), **P**art (**P**) or **B**ody (**B**) scale graph modeling with J, 10, and 5 nodes.

Evaluation on Different-Scale Skeleton Graphs. We construct different-scale graphs (Rao et al. 2021a) for MoCos learning to evaluate its performance under varying graph modeling. As presented in Table 4, compared with the state-of-the-art multi-scale graph method SM-SGE (Rao et al. 2021a), our model achieves better performance in most cases of both original and higher level skeleton representations ($e.g.$, part-scale skeleton graphs). Such results suggest the compatibility of the proposed motif guided graph transformer (MGT) with different-scale graph modeling, and also justify its stronger capability to learn more effective graph features and semantics at different levels for person re-ID.

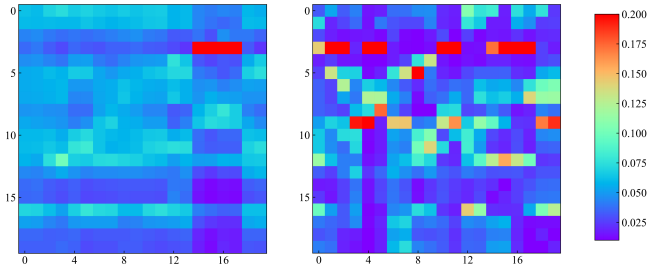
Transfer to Unsupervised Paradigms. As shown in Table 5, our MGT and CSP (denoted as "+ MoCos") can be transferred for *unlabeled* skeleton relation and prototype learning, which effectively boosts performance of different unsupervised non-graph and non-transformer models (Rao and Miao 2022; Rao and Miao 2022) in most cases. This

Methods	BIWI-S		BIWI-W		KS20		KGBD	
	mAP	R_1	mAP	R_1	mAP	R_1	mAP	R_1
SPC-MGR (2022)	16.0	34.1	19.4	18.9	21.7	59.0	6.9	40.8
SPC-MGR + MoCos	**16.3**	**42.8**	**20.1**	**23.6**	**23.6**	**65.4**	**8.2**	**43.2**
SimMC (2022)	12.3	41.7	19.9	24.5	22.3	66.4	11.7	**54.9**
SimMC + MoCos	**16.0**	**55.4**	**22.4**	**25.9**	**23.9**	**67.2**	**12.1**	54.6

Table 5: Performance of our approach when applied to different unsupervised paradigms using *unlabeled* skeletons.



(a) t-SNE Visualization



(b) Relation Comparison

Figure 3: (a) t-SNE visualization of features for the first ten classes in IAS and KS20. Different colors indicates different classes. (b) Visualization of mean relation values inferred by non-motif method (Rao and Miao 2023) (Left) and our MoCos (Right) on the same value scale and testing skeletons.

validates generality and scalability of MoCos, which can be potentially applied to more general scenarios without labels.

Feature and Relation Visualization. The t-SNE visualization (Van der Maaten and Hinton 2008) in Fig. 3 (a) shows the evident inter-class separation of the learned features, suggesting the effectiveness of our approach to capture useful class-related semantics on different datasets. We also visualize the mean relations of ($J = 20$) joints inferred from MoCos in Fig. 3 (b), and the results imply that our motif-guided approach could capture richer and more salient joint correlations than TranSG (Rao and Miao 2023) that solely uses full-relation learning without motifs. More empirical and theoretical analyses are provided in Appendix I and II.

Conclusion

In this paper, we propose MoCos to perform motif-guided joint relation learning and combinatorial skeleton prototype learning for person re-ID. We design the motif guided graph transformer (MGT) that incorporates hierarchical structural motifs and gait collaborative motifs to capture key relations within multi-order body joints' structure and gait-related limbs. The combinatorial skeleton prototype learning (CSP) is proposed to contrast randomly-combined sub-skeleton and sub-tracklet graph features with skeleton prototypes to learn class-related semantics and discriminative representations. Our approach outperforms existing state-of-the-art models, and can be generally applied to various scenarios.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD/2022-01-034[T]).

References

Andersson, V. O.; and Araujo, R. M. 2015. Person identification using anthropometric and gait data from Kinect sensor. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 425–431.

Barbosa, I. B.; Cristani, M.; Del Bue, A.; Bazzani, L.; and Murino, V. 2012. Re-identification with RGB-D sensors. In *European Conference on Computer Vision (ECCV) Workshop*, 433–442. Springer.

Boluki, S.; Ardywibowo, R.; Dadaneh, S. Z.; Zhou, M.; and Qian, X. 2020. Learnable Bernoulli dropout for Bayesian deep learning. In *International Conference on Artificial Intelligence and Statistics*, 3905–3916. PMLR.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2019. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1): 172–186.

Chen, C.-H.; and Ramanan, D. 2017. 3D human pose estimation= 2D pose estimation+ matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7035–7043.

Davis, J. V.; Kulis, B.; Jain, P.; Sra, S.; and Dhillon, I. S. 2007. Information-theoretic metric learning. In *International Conference on Machine Learning (ICML)*, 209–216.

Dwivedi, V. P.; and Bresson, X. 2021. A generalization of transformer networks to graphs. In *AAAI Conference on Artificial Intelligence (AAAI) Workshop*.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the ACM International Conference on Multimedia*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; and Cristani, M. 2010. Person re-identification by symmetry-driven accumulation of local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2360–2367. IEEE.

Feng, S.; Miao, C.; Zhang, Z.; and Zhao, P. 2024. Latent diffusion transformer for probabilistic time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 11979–11987.

Gao, Z.; Jiang, C.; Zhang, J.; Jiang, X.; Li, L.; Zhao, P.; Yang, H.; Huang, Y.; and Li, J. 2023. Hierarchical graph learning for protein–protein interaction. *Nature Communications*, 14(1): 1093.

Gray, D.; and Tao, H. 2008. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 262–275. Springer.

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

Hasan, M.; and Babaguchi, N. 2016. Long-term people re-identification using anthropometric signature. In *International Conference on Biometrics Theory, Applications and Systems*, 1–6. IEEE.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 3316–3333.

Liao, R.; Yu, S.; An, W.; and Huang, Y. 2020. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98: 107069.

Liu, Z.; Zhang, Z.; Wu, Q.; and Wang, Y. 2015. Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168: 1144–1156.

Lu, J.; Wan, H.; Li, P.; Zhao, X.; Ma, N.; and Gao, Y. 2023. Exploring High-order Spatio-temporal Correlations from Skeleton for Person Re-identification. *IEEE Transactions on Image Processing*.

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

Munaro, M.; Fossati, A.; Basso, A.; Menegatti, E.; and Van Gool, L. 2014b. One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*, 161–181. Springer.

Munaro, M.; Ghidoni, S.; Dizmen, D. T.; and Menegatti, E. 2014c. A feature-based approach to people re-identification using skeleton keypoints. In *International Conference on Robotics and Automation (ICRA)*, 5644–5651. IEEE.

Murray, M. P.; Drought, A. B.; and Kory, R. C. 1964. Walking patterns of normal men. *Journal of Bone and Joint Surgery*, 46(2): 335–360.

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

Nguyen, V. D.; Mirza, S.; Mantini, P.; and Shah, S. K. 2024. Attention-based shape and gait representations learning for video-based cloth-changing person re-identification. *arXiv preprint arXiv:2402.03716*.

Pala, P.; Seidenari, L.; Berretti, S.; and Del Bimbo, A. 2019. Enhanced skeleton and face 3D data for person re-identification from depth cameras. *Computers & Graphics*, 79: 69–80.

Pržulj, N. 2007. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2): e177–e183.

Rao, H.; Hu, X.; Cheng, J.; and Hu, B. 2021a. 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*, 1812–1820.

Rao, H.; Leung, C.; and Miao, C. 2024. Hierarchical skeleton meta-prototype contrastive learning with hard skeleton mining for unsupervised person re-identification. *International Journal of Computer Vision*, 132(1): 238–260.

Rao, H.; Li, Y.; and Miao, C. 2022. Revisiting k-Reciprocal Distance Re-Ranking for Skeleton-Based Person Re-Identification. *IEEE Signal Processing Letters*, 29: 2103–2107.

Rao, H.; and Miao, C. 2022. SimMC: Simple Masked Contrastive Learning of Skeleton Representations for Unsupervised Person Re-Identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 1290–1297.

Rao, H.; and Miao, C. 2022. Skeleton Prototype Contrastive Learning with Multi-Level Graph Relation Modeling for Unsupervised Person Re-Identification. *arXiv preprint arXiv:2208.11814*.

Rao, H.; and Miao, C. 2023. TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning with Structure-Trajectory Prompted Reconstruction for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rao, H.; and Miao, C. 2024. A Survey on 3D Skeleton Based Person Re-Identification: Approaches, Designs, Challenges, and Future Directions. *arXiv preprint arXiv:2401.15296*.

Rao, H.; Wang, S.; Hu, X.; Tan, M.; Da, H.; Cheng, J.; and Hu, B. 2020. Self-Supervised Gait Encoding with Locality-Aware Attention for Person Re-Identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, 898–905.

Rao, H.; Wang, S.; Hu, X.; Tan, M.; Guo, Y.; Cheng, J.; Liu, X.; and Hu, B. 2021b. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6649–6666.

Rao, H.; Xu, S.; Hu, X.; Cheng, J.; and Hu, B. 2021c. Multi-Level Graph Encoding with Structural-Collaborative Relation Learning for Skeleton-Based Person Re-Identification. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 973–980.

Rao, H.; Zeng, M.; Zhao, X.; and Miao, C. 2024. A Survey of Artificial Intelligence in Gait-Based Neurodegenerative Disease Diagnosis. *arXiv preprint arXiv:2405.13082*.

Sankar, A.; Zhang, X.; and Chang, K. C.-C. 2017. Motif-based convolutional neural network on graphs. *arXiv preprint arXiv:1711.05697*.

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

Sporns, O.; and Kötter, R. 2004. Motifs in brain networks. *PLoS biology*, 2(11): e369.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11): 2579–2605.

Vezzani, R.; Baltieri, D.; and Cucchiara, R. 2013. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 46(2): 1–37.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2016. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12): 2501–2514.

Wang, Z.; Wei, D.; Hu, X.; and Luo, Y. 2020. Human skeleton mutual learning for person re-identification. *Neurocomputing*, 388: 309–323.

Weinberger, K. Q.; and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(2): 207–244.

Wen, Y.-H.; Gao, L.; Fu, H.; Zhang, F.-L.; and Xia, S. 2019. Graph CNNs with motif and variable temporal block for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8989–8996.

Wen, Y.-H.; Gao, L.; Fu, H.; Zhang, F.-L.; Xia, S.; and Liu, Y.-J. 2022. Motif-GCNs with local and non-local temporal blocks for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2009–2023.

Wu, A.; Zheng, W.-S.; and Lai, J.-H. 2017. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6): 2588–2603.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 7444–7452.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6): 2872–2893.

Yu, S.; Tan, D.; and Tan, T. 2006. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International Conference on Pattern Recognition (ICPR)*, volume 4, 441–444. IEEE.

Zhang, P.; Xu, J.; Wu, Q.; Huang, Y.; and Ben, X. 2020. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE Transactions on Multimedia*, 23: 3562–3576.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1116–1124.

Zhicheng, C.; Shibo, F.; Zhang, Z.; Xiao, X.; Gao, X.; and Zhao, P. 2024. SDformer: Similarity-driven Discrete Transformer For Time Series Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.