Path-JEPA: Path Signature Based Predictive Learning for Skeleton Action Recognition

Ashutosh Singh[1] and Ashish Singh[2]

[1] Northeastern University, Boston, MA 02115, USA
singh.ashu@northeastern.edu
[2] University of Massachusetts Amherst, Amherst, MA 01003, USA
ashishsingh@cs.umass.edu

Abstract. Self-supervised learning (SSL) has become a leading paradigm for skeleton-based action recognition, yet the choice of prediction target remains a central limitation. Existing masked modeling methods typically reconstruct raw joint coordinates, emphasizing low-level detail and remaining sensitive to noise, while recent non-reconstruction variants predict learned latent features that improve semantics but capture little explicit motion structure. We propose **Path-JEPA**, a self-supervised predictive learning framework that adopts path signatures as the prediction target for skeleton sequences. Rather than predicting coordinates or abstract embeddings, Path-JEPA predicts latent representations of continuous-time, multi-scale geometric descriptors computed over joint, edge, and chain paths, capturing motion across multiple spatial and temporal scales. These targets encode displacement, signed area, and higher-order interactions, yielding a geometry-grounded representation of human motion that is naturally robust to temporal resampling and frame-rate variation. To make such targets effective within a JEPA framework, we introduce *signature-augmented masking*, which propagates joint-space masks to all dependent signature tokens, forcing the model to infer the missing motion geometry from broader anatomical and temporal context. Because signatures summarize motion over intervals rather than individual frames, they also yield greater computational efficiency on long sequences. Extensive experiments on NTU RGB+D 60, NTU RGB+D 120, and PKU-MMD show that Path-JEPA learns stronger representations than prior masked prediction methods, achieving state-of-the-art performance across diverse downstream tasks while exhibiting improved robustness to irregular sampling. Project website:Path-JEPA.

1 Introduction

Human action recognition is central to human-robot interaction [31, 50], intelligent surveillance [41, 49], and behavior analysis [2, 42, 46]. Skeleton sequences provide a compact, privacy-preserving representation of human action while retaining kinematic structure [8, 11, 15]. Self-supervised learning has emerged as the dominant paradigm for learning such representations, with recent work shifting from contrastive objectives [13, 18, 22, 25, 28] toward masked predictive

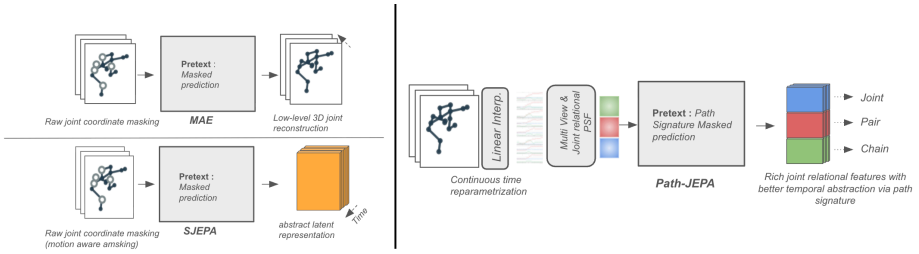

Fig. 1: Left: MAE-style pretext task masks raw 3D joint coordinates and reconstructs them, producing low-level spatial features tied closely to input coordinates. **Bottom Left**: S-JEPA also masks in joint-coordinate space but predicts high-level latent targets, yielding stronger representations while still ignoring explicit spatio-temporal structure. **Right**: Path-JEPA first lifts skeleton sequences into rich multiscale path-signature (PSF) tokens over joint, edge, and chain paths, then performs JEPA-style masked prediction in this signature space, encouraging richer, geometrically grounded spatio-temporal representations.

learning—a scalable approach that avoids augmentation engineering and pairs naturally with transformer architectures [14, 19].

The effectiveness of masked skeleton SSL depends critically on the prediction target. Coordinate-reconstruction methods [37, 52, 58] predict raw joint coordinates or motion-derived quantities, keeping the pretext task close to low-level signal recovery and making it sensitive to noise and temporal discretization. More recent work explores richer targets [48, 55, 56]. JEPA-style methods [4–6, 23, 39] go further by predicting in latent space; S-JEPA [1] and GFP [48] show that representation-space prediction substantially outperforms coordinate recovery for skeleton sequences. Yet all these methods share a deeper limitation: their targets are derived from masked raw joints and do not explicitly encode the *relational geometry* of motion—the spatiotemporal coupling between joints and body parts. Action semantics depend not on isolated joint movement but on coordinated patterns across neighboring and distant joints over multiple temporal scales [9, 12, 32, 53]. Predicting targets that lack this structure forces the model to recover it only implicitly.

A second limitation is that human motion is inherently continuous, yet models operate on discrete frame-indexed observations, creating frame-rate dependencies that hinder generalization under resampling. A principled solution requires targets that abstract motion over intervals rather than individual timesteps.

Path signatures address both limitations simultaneously. Originating from rough path theory [7, 17, 34, 36], they represent a trajectory through iterated integrals, providing principled temporal abstraction over intervals. By lifting discrete frames to continuous paths via linear interpolation, they summarize motion geometry in a form robust to frame-rate variation and irregular sampling. When computed over anatomically meaningful multi-joint paths—individual joints, joint pairs, and kinematic chains—they further capture higher-order spatiotemporal coupling across the body [10, 54]. Prior work has established the value of

path signatures for supervised gesture and action recognition [10, 20, 24, 26, 54]; to our knowledge they have not yet been used as prediction targets in a masked SSL paradigm.

We propose **Path-JEPA**, a self-supervised framework that predicts masked targets derived from *path-signature representations* of skeleton motion. We lift each skeleton sequence to a hierarchy of multi-scale signature tokens over joint, pair, and kinematic-chain paths, providing relational and temporally abstracted targets consistent with the JEPA principle. We introduce **signature-augmented masking**, which propagates joint masking to all dependent signature tokens, enforcing structured prediction over motion geometry. Our encoders are transformers that operate directly over signature tokens [40], enabling richer spatial reasoning over relational structures than raw-joint tokenization affords. We present extensive experimentation across different benchmarks, baselines and downstream task. We observe that Path-JEPA performs well across all evaluations beating SOTA in most cases. Our contributions are:

- We introduce **Path-JEPA**, a JEPA-based SSL framework for skeleton action recognition that predicts embeddings derived from path-signature features rather than raw coordinates.
- We propose **multi-scale relational signature tokenization** over joint, pair, and kinematic-chain paths, capturing relational structure and temporal abstraction over continuously interpolated motion.
- We introduce **signature-augmented masking**, which propagates joint masking to all dependent signature tokens, enforcing structured prediction over motion geometry.
- We demonstrate improved linear evaluation, fine-tuning, semi-supervised, and cross-dataset performance, with enhanced robustness to temporal resampling and irregular sampling.

2 Related Work

2.1 Self-Supervised Skeleton Action Recognition

Self-supervised methods for skeleton action recognition fall into three paradigms. **Contrastive methods** [13,18,22,25,28] learn view-invariant representations by contrasting augmented sequence pairs, but rely heavily on hand-crafted augmentation strategies and are susceptible to false negatives. Beyond single-stream skeleton contrastive learning, several methods have explored richer sources of supervision and representation design. Sun et al. [47] learn unified multi-modal unsupervised representations for skeleton-based action understanding by leveraging complementary skeleton, RGB, and text modalities. Lin et al. [27] introduce an idempotent unsupervised learning objective for skeleton action recognition in coordinate space, while Wang et al. [51] study heterogeneous skeleton representations for more flexible action representation learning. **Reconstruction-based methods** train models to recover corrupted or masked inputs. LongT-

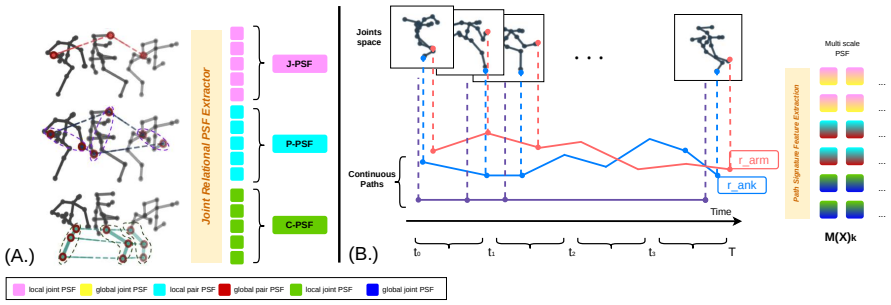

Fig. 2: Path-signature feature extraction for skeleton sequences. **(a)** Spatial path construction. We define three families of paths: joint paths (J-PSF, pink) following individual joint trajectories, pairwise/edge paths (P-PSF, blue) following relative motion between kinematically connected joints, and chain paths (C-PSF, teal) following multi-joint limb chains. Each path is mapped to a path-signature feature (PSF). **(b)** Temporal windowing and multi-view signatures. The sequence is partitioned into (K) windows ($[t_{k-1},\ t_k]$). For every path and window, we compute a PSF that concatenates a global prefix log-signature ($S_{0,T}$) and a local window log-signature (S_{t_{k-1},t_k}). Stacking the J-PSF, P-PSF, and C-PSF across all paths and windows yields the token matrix ($M(X)_k$), which serves as the input to the Path-JEPA encoder.

GAN [57] reconstructs masked joints via adversarial training; P&C [45] predicts future frames as the learning objective. SkeletonMAE [52] applies graph-based masked autoencoding at a 90% masking ratio, and MAMP [37] extends this with motion-aware masking that prioritizes high-movement joints. MotionBERT [58] uses spatio-temporal transformers for a 2D-to-3D lifting pretext task. Subsequent work explores diffusion-based targets [56], richer feature reconstruction [55], and asymmetric masking strategies [3, 38]. Despite these advances, reconstruction-based methods focus on low-level coordinate recovery, limiting semantic expressiveness and robustness to temporal variation. **Representation prediction methods** address this by predicting in the latent space. S-JEPA [1] introduced JEPA-style prediction for skeletons, training a student encoder to predict teacher latent representations for masked joints using cross-entropy loss with centering for stability. GFP [48] extends this with hierarchical feature prediction across temporal granularities and variance-covariance regularization to prevent collapse. Both methods substantially outperform coordinate reconstruction, but their targets are still derived from masked raw joints and lack explicit encoding of spatiotemporal relational geometry or mathematical guarantees of temporal invariance, which are precisely the limitations Path-JEPA addresses.

2.2 Path Signatures for Action Recognition

Path signatures have proven effective in supervised skeleton recognition. Yang et al. [54] introduced signatures for landmark-based action recognition using spatial and temporal joint paths with a lightweight classifier. Li et al. [24] develop

a dual-stream spatio-temporal signature block, and LogSig-RNN [26] combines log-signatures with recurrent networks for robust action recognition. Jiang et al. [20] combine GCN-based spatial encoding with signature LSTM modeling. Cheng et al. [10] propose learnable path generation via self-attention to adaptively select semantically relevant joints. The Rough Transformer [40] integrates signature patching into transformer architectures for continuous time-series modeling. While these methods establish the value of signature features, all operate in a supervised setting and treat signatures as fixed input features rather than structured prediction targets. Path-JEPA is the first to integrate path signatures into a masked predictive SSL framework, using multi-scale signature tokenization and signature-augmented masking to provide principled, temporally invariant prediction targets.

3 Method

Path-JEPA learns skeleton action representations via self-supervised masked prediction over *signature tokens* (Fig. 3). Given a skeleton sequence $X \in \mathbb{R}^{T\times J\times 3}$ with T frames and J joints, we (i) lift discrete joint trajectories into continuous paths, (ii) compute multi-scale *multi-view* (global+local) log-signature descriptors over three spatial modalities (joint, edge, chain), and (iii) train a JEPA framework to predict masked signature-token embeddings. Concretely, a student encoder f_θ processes only visible tokens, a lightweight predictor g_ϕ reconstructs masked token embeddings via cross-attention, and an EMA teacher encoder $f_{\bar{\theta}}$ provides targets. Compared to coordinate recovery or predicting teacher embeddings of joint tokens, Path-JEPA learns from temporally abstract, relationally structured motion descriptors, while masking is propagated through the token hierarchy to prevent leakage.

3.1 Path Signature Representation

Continuous paths from discrete skeletons. For each joint $j \in \{1,\dots,J\}$ with discrete coordinates $X_j \in \mathbb{R}^{T\times 3}$, we construct a continuous-time path by piecewise-linear interpolation $\tilde{X}_j : [0,\tau] \to \mathbb{R}^3$, where τ is the sequence duration. To encode temporal variation more richly, we apply a lead–lag lift by augmenting each path with a time-delayed copy:

$$\tilde{X}_j^{\mathrm{LL}}(t) = [\tilde{X}_j(t), \tilde{X}_j(t-\delta)] \in \mathbb{R}^{d'}, \qquad d' = 6, \tag{1}$$

so differences between the two streams implicitly encode velocity-like information.

The truncated signature. For a path $\mathbf{x} : [s,t] \to \mathbb{R}^d$, the truncated signature up to level n is

$$S(\mathbf{x})_{s,t}^{\le n} = \left(1, S^{(1)}, S^{(2)}, \dots, S^{(n)}\right), \tag{2}$$

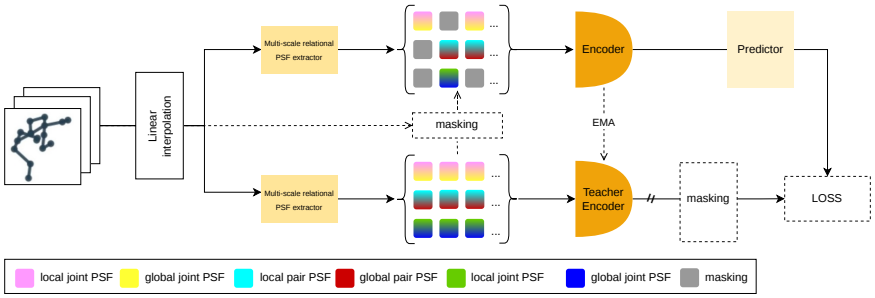

Fig. 3: Overview of Path-JEPA. The input skeleton sequence X is first mapped to path-signature features (PSF) over joint, edge, and chain paths. A joint-space mask is then propagated to PSF tokens, producing masked tokens for the context encoder (f_θ) (top) and a shared index set ($\mathcal{I}$) used to select targets from the EMA teacher encoder ($f_{\bar{\theta}}$) (bottom), which always sees the full, unmasked PSF sequence. The predictor (g_ϕ) operates on the context encoder output and is trained to reconstruct teacher features at masked indices, using a cosine JEPA loss ($\mathcal{L}_{\text{JEPA}}$) defined only on ($\mathcal{I}$).

where $S^{(k)}$ contains d^k iterated-integral coefficients. The level-1 term encodes displacement $S_i^{(1)} = \int_s^t d\mathbf{x}^i = \mathbf{x}^i(t) - \mathbf{x}^i(s)$. Level-2 terms encode signed areas via $S_{ij}^{(2)} = \int_s^t \int_s^{u_1} d\mathbf{x}^i \, d\mathbf{x}^j$, capturing Lévy areas when $i \neq j$.

Log-signature compression. After the lead–lag lift to $\mathbb{R}^{d'}$, the uncompressed truncated signature has $D_{\text{sig}} = \sum_{k=0}^{n} (d')^k$ coefficients. We use log-signatures to reduce redundancy (via the BCH structure [16, 21, 35]), yielding $D_{\log}$ coefficients. For example, with $d' = 6$ and $n = 2$, $D_{\text{sig}} = 43$ while our log-signature representation has $D_{\log} = 22$ coefficients. Signatures also inherit useful continuous-time properties from rough path theory [7,17,34,36]; in particular, they summarize trajectories over intervals rather than individual frames, improving robustness to temporal discretization (detailed description of path signature background in **appendix A**).

3.2 Multi-Scale Signature Tokenization

A core design goal of Path-JEPA is to construct prediction targets that are both *relationally structured* and *temporally abstracted*. To this end, we build a set of signature tokens over (i) a hierarchy of temporal windows at multiple scales, and (ii) a kinematic spatial hierarchy that spans individual joints, **joint pairs**, and multi-joint anatomical chains (Fig. 2). Each token is a deterministic, mathematically principled descriptor of continuous motion geometry over a specific path instance and temporal interval, rather than a raw frame-level observation.

Temporal Multi-View Tokens (Global + Local) A single temporal window captures only a fragment of the motion; a single global summary loses

local dynamics. We therefore design each token to simultaneously encode both short-range and long-range temporal structure by combining a *global* and a *local* signature view of the *same underlying path instance*.

Concretely, we build a hierarchy of temporal scales indexed by $\ell \in \{0, \dots, L-1\}$, where scale ℓ uses windows of length $W_\ell = 2^\ell W_0$ with stride s_ℓ. This produces a set of non-overlapping (or strided) temporal windows $\omega_{\ell,k} = (t_{\ell,k-1}, t_{\ell,k}]$ at each scale. Coarser scales capture slow, global motion trends, while finer scales capture fast, local dynamics—together spanning the full range of temporal structure relevant to action recognition.

Let p denote a path instance (a joint trajectory, a **joint-pair** displacement, or a chain concatenation) with associated continuous lead–lag trajectory $\tilde{X}_p^{\mathrm{LL}}$. The lead–lag transformation lifts the discrete trajectory into a continuous two-channel path, enabling the log-signature to capture both the path's value and its direction of travel at each moment. For each window $\omega_{\ell,k}$, we define the *multi-view token* as the concatenation of a *global* log-signature computed over the **entire sequence** and a *local* log-signature computed over the current window:

$$m_{p,\ell,k} \;=\; \Big[\; \log S(\tilde{X}_p^{\mathrm{LL}})_{0,T} \;,\; \log S(\tilde{X}_p^{\mathrm{LL}})_{t_{\ell,k-1},t_{\ell,k}} \;\Big] \in \mathbb{R}^{D_{\mathrm{mv}}}, \quad D_{\mathrm{mv}} = 2D_{\log} \tag{3}$$

The first term is the *global view*: a log-signature over $[0, T]$ that provides a sequence-level temporal abstraction of the full motion trajectory for path instance p. The second term is the *local view*: a log-signature restricted to the current window $\omega_{\ell,k}$, capturing fine-grained dynamics within that interval. By concatenating both views into a single token, each $m_{p,\ell,k}$ carries complementary information about the overall motion pattern and the local temporal context, without requiring separate streams or additional fusion modules.

Spatial Kinematic Hierarchy (Joint / Pair / Chain Paths) To encode relational structure across the body, we define three levels of spatial path instances that mirror the natural kinematic organization of the skeleton:

$$\mathcal{P}_J = \{1, \dots, J\}, \quad \mathcal{P}_P = \mathcal{E} \text{ (kinematic joint pairs)}, \quad \mathcal{P}_C = \mathcal{C} \text{ (predefined chains)}$$

Each level captures motion at a different spatial granularity, from individual joints to multi-joint coordinated patterns.

Joint paths. The most fine-grained level treats each joint $j \in \mathcal{P}_J$ independently. We apply the lead–lag transform to the raw joint trajectory to obtain $\tilde{X}_j^{\mathrm{LL}}$, and form tokens $m_{j,\ell,k}$ via Eq. 3. These tokens encode the intrinsic motion of each joint in isolation, providing localized descriptors of individual joint dynamics.

Joint-pair paths. The intermediate level captures pairwise relational motion between anatomically connected joints. For each kinematic pair $(u, v) \in \mathcal{P}_P$, we define a relative displacement path

$$Y_{u,v}(t) = \tilde{X}_v(t) - \tilde{X}_u(t),$$

which describes how joint v moves relative to joint u. We apply the lead–lag transform to obtain $\tilde{Y}^{\mathrm{LL}}_{u,v}$ and form tokens $m_{(u,v),\ell,k}$ via Eq. 3. Using relative displacements rather than absolute coordinates makes these tokens naturally invariant to global body translation. For a tree-structured skeleton, $|\mathcal{P}_P| = J-1$, so this level provides nearly the same coverage as joint tokens but with explicit pairwise relational encoding.

Chain paths. The coarsest spatial level captures coordinated multi-joint motion along anatomically meaningful limb segments. For each predefined chain $c \in \mathcal{P}_C$ (e.g., shoulder→elbow→wrist, or hip→knee→ankle), we construct a chain path Z_c by concatenating the constituent joint-pair displacement trajectories in anatomical order. We apply the lead–lag transform to obtain $\tilde{Z}^{\mathrm{LL}}_c$ and form tokens $m_{c,\ell,k}$ via Eq. 3. Chain tokens are particularly important for capturing the higher-order joint coupling that underlies action semantics—a reaching action, for instance, is better characterized by the coordinated displacement along the arm chain than by any individual joint pair in isolation.

Token set per window. At each temporal scale and window index (ℓ, k), we collect tokens over all path instances across the three spatial levels:

$$M(X)_{\ell,k} \;=\; \{\, m_{p,\ell,k} : p \in \mathcal{P}_J \cup \mathcal{P}_P \cup \mathcal{P}_C \,\}. \tag{4}$$

The full token sequence fed to the transformer is the union over all (ℓ, k), ordered by temporal scale and time. This unified token set allows the transformer to attend jointly over joint, joint-pair, and chain tokens at multiple temporal resolutions within a single sequence, enabling rich cross-level and cross-scale reasoning without requiring separate processing streams.

Projection and positional embeddings. Raw signature tokens have dimension D_{mv} and vary in their geometric meaning depending on whether they encode a joint, joint-pair, or chain path. We therefore project each token to the transformer dimension d_{model} using *type-specific* linear projections:

$$\tilde{m}_{p,\ell,k} = W_{\mathrm{type}(p)}\, m_{p,\ell,k} + e_{\mathrm{type}(p)} + e_{\mathrm{struct}}(\ell, d_G(p)), \tag{5}$$

where $W_{\mathrm{type}(p)} \in \mathbb{R}^{d_{\mathrm{model}} \times D_{\mathrm{mv}}}$ is a learned projection specific to the token type (joint, joint-pair, or chain), and $e_{\mathrm{type}(p)}$ is a learned type embedding that informs the transformer of the spatial level each token comes from. The structural embedding $e_{\mathrm{struct}}(\ell, d_G(p))$ jointly encodes the temporal scale ℓ and the kinematic depth $d_G(p)$—the graph distance from the skeleton root to the joint or joints associated with path instance p—providing the transformer with a coherent spatial-temporal coordinate system for attending over tokens from different levels and scales. Together, these embeddings allow the model to reason about both *what* a token represents spatially and *when* it occurs temporally.

3.3 Spatial Processing via Signature Aggregation

Our tokenization makes relational structure explicit: edge tokens encode pairwise joint relationships through relative-motion signatures, and chain tokens aggregate these relations along anatomical limbs. The student encoder processes all

tokens jointly with self-attention. We optionally augment attention logits with learned bias terms that encode temporal and kinematic structure:

$$\alpha_{ij} = \frac{Q_i K_j^\top}{\sqrt{d_k}} + \beta_t \kappa_t(|t_i - t_j|) + \beta_\ell \kappa_\ell(\ell_i, \ell_j) + \beta_\omega \mathrm{IoU}(\omega_i, \omega_j) + \beta_G \kappa_G(d_G(i,j)), \quad (6)$$

where token i has metadata (center time t_i, scale ℓ_i, window ω_i), and $d_G(i,j)$ is the kinematic graph distance between the joints associated with tokens i and j (e.g., nearest endpoint joints for edges, endpoints for chains). Scalars $\{\beta_t, \beta_\ell, \beta_\omega, \beta_G\}$ are learned.

3.4 Signature-Augmented Masking

Masking is applied *after* signature extraction: we compute all tokens $M(X)_{\ell,k}$ from the full sequence, then mask tokens in the student/context branch (we never zero coordinates before computing signatures).

Token dependency masking. Let $\mathcal{M} \subset \{1, \ldots, J\}$ denote the masked joint set. We mask every token whose underlying path touches any masked joint:

$$\mathcal{I}(\mathcal{M}) \;=\; \{\, (p, \ell, k) \;:\; p \cap \mathcal{M} \neq \emptyset \,\}. \quad (7)$$

Equivalently: (i) joint tokens for $j \in \mathcal{M}$, (ii) edge tokens with an endpoint in $\mathcal{M}$, and (iii) chain tokens traversing any joint in $\mathcal{M}$. Because each token is the concatenation $[\text{global}, \text{local}]$ for the *same* path (Eq. 3), masking a token removes *both* its global and local views. This prevents information leakage through overlapping relational paths and enforces leak-free prediction.

Kinematically coherent motion-aware sampling. In practice, we build $\mathcal{M}$ via a hybrid strategy: we compute per-joint motion magnitude, sample a seed joint with higher probability for high-motion joints, then expand the mask to the seed joint's kinematic subtree (e.g., an entire limb branch), enforcing anatomical coherence while retaining motion awareness.

3.5 JEPA Training Framework

Figure 3 shows the training workflow.

Student encoder. The student encoder f_θ is a transformer with L_e layers. It processes only the *context* tokens

$$\mathcal{C} = \{\, \tilde{m}_{p,\ell,k} : (p, \ell, k) \notin \mathcal{I}(\mathcal{M}) \,\},$$

and outputs contextual embeddings $Z_\mathcal{C} \in \mathbb{R}^{|\mathcal{C}| \times d_{\text{model}}}$ using attention in Eq. 6.

Predictor. For each masked token index $(p, \ell, k) \in \mathcal{I}(\mathcal{M})$, we construct a learnable query embedding $q_{p,\ell,k}$ from metadata (token type, path ID, time center, scale, and kinematic depth). The predictor g_ϕ (with L_p layers) performs cross-attention from these queries to the student context:

$$\hat{Z}_\mathcal{I} = g_\phi(Q_\mathcal{I}, Z_\mathcal{C}), \quad (8)$$

producing predictions $\hat{Z}_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}| \times d_{\text{model}}}$ aligned with the masked indices $\mathcal{I} = \mathcal{I}(\mathcal{M})$.

Teacher encoder. The teacher encoder $f_{\bar{\theta}}$ has the same architecture as the student and processes the full (unmasked) token sequence to produce $Z_{\text{full}} \in \mathbb{R}^{N_{\text{total}} \times d_{\text{model}}}$. Targets are extracted at masked indices:

$$Z^*_{\mathcal{I}} = Z_{\text{full}}[\mathcal{I}] \in \mathbb{R}^{|\mathcal{I}| \times d_{\text{model}}}. \tag{9}$$

Teacher parameters are updated by EMA: $\bar{\theta} \leftarrow \lambda\bar{\theta} + (1-\lambda)\theta$. We apply standard centering for stability: $Z^*_{\mathcal{I}} \leftarrow Z^*_{\mathcal{I}} - c$, where c is a running mean with decay β.

Objective. We minimize cosine distance between predicted and target embeddings:

$$\mathcal{L}_{\text{JEPA}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(1 - \frac{\hat{\mathbf{z}}_i^\top \mathbf{z}_i^*}{\|\hat{\mathbf{z}}_i\| \, \|\mathbf{z}_i^*\|} \right). \tag{10}$$

After pretraining, the teacher encoder $f_{\bar{\theta}}$ is used as the feature extractor for linear probing and fine-tuning on downstream action recognition.

4 Experiments

Datasets: We evaluate Path-JEPA on four widely-used skeleton-based action recognition benchmarks. NTU RGB+D 60 (NTU60) [43]contains 56,880 action sequences with 60 classes, evaluated on Cross-Subject (X-Sub) and Cross-View (X-View) protocols. NTU RGB+D 120 (NTU120) [30] extends this to 114,480 sequences across 120 classes, with Cross-Subject (X-Sub) and Cross-Setup (X-Set) protocols. PKU-MMD [29] provides 1,076 untrimmed sequences across Phase I (51 classes) and Phase II (41 classes) following cross-subject evaluation.

Implementation Details: Our architecture uses vanilla transformer backbones [14] with $L_e = 6$ encoder layers and $L_p = 4$ predictor layers, both with $d_{\text{model}} = 512$, 8 attention heads, and feedforward dimension 2048. We pretrain for 300 epochs using AdamW [33] with learning rate 3×10^{-4}, weight decay 0.05, and batch size 256, applying cosine decay with 10-epoch warmup. Teacher EMA momentum increases from 0.996 to 1.0 via cosine schedule, with centering momentum $\beta = 0.9$. Signatures use truncation level $n = 2$ with lead-lag transformation ($d' = 6$), yielding $D_{\log} = 22$ coefficients per token. Multi-scale windows employ $\ell = 0$ (size 25, stride 12) and $\ell = 1$ (size 50), resulting in $L_{\text{total}} = 4$ windows and $N_{\text{total}} = 152$ tokens per sequence. For downstream evaluation, linear probing trains for 100 epochs with SGD (momentum 0.9, learning rate 0.1), while fine-tuning uses 300 epochs with AdamW (learning rate 3×10^{-4}, layer-wise decay 0.75). We provide additional experiments namely - computational comparison, sensitivity to β parameters, frame drop and qualitative results in Appendix B.

4.1 Linear Evaluation

Table 1 presents linear evaluation results where the pretrained teacher encoder is frozen and only a linear classifier is trained. Path-JEPA achieves 87.6% on

Table 1: Linear evaluation accuracy (%) on skeleton action recognition benchmarks. Best results are in **bold**, and second best are underlined.

Method	In.	NTU60		NTU120		PKU-MMD	
		X-Sub	X-View	X-Sub	X-Set	Ph-I	Ph-II
Contrastive learning							
CrosSCLR [25]	J+B+M	77.8	83.4	67.9	66.7	84.9	21.2
AimCLR [18]	J+B+M	78.9	83.8	68.2	68.8	87.4	39.5
ActCLR [28]	J+B+M	84.3	88.8	74.3	75.7	–	–
HiCo [13]	J	81.1	88.6	72.8	74.1	89.3	49.4
Coordinate reconstruction							
SkeletonMAE [52]	J	74.8	77.7	72.5	73.5	82.8	36.1
MAMP [37]	J	84.9	89.1	78.6	79.1	92.2	53.8
Feature prediction							
GFP [48]	J	85.9	92.0	79.1	80.3	–	56.2
S-JEPA [1]	J	85.3	89.8	79.6	79.9	92.2	53.5
Path-JEPA	J	**87.6**	**92.9**	**81.5**	**81.8**	**93.4**	**56.7**

NTU60 X-Sub, surpassing GFP by 1.7% and (ActCLR) by 3.3%. On NTU120, Path-JEPA reaches 81.5% on X-Sub and 81.8% on X-Set, improving over S-JEPA by 1.9% and GFP by 1.5%. On PKU-MMD, Path-JEPA achieves 93.4% on Phase I and 56.7% on Phase II, matching or exceeding previous best results. These improvements across diverse datasets show that signature-based representations provide superior features compared to learned embeddings or coordinate-level representations. The geometric properties encoded by path signatures, displacement, Lévy area, and higher-order couplings, prove more informative than raw coordinates or black-box learned features for action recognition.

4.2 Fine-tuning Evaluation

Table 2 presents results under the fine-tuning protocol where the pretrained encoder is adapted to the downstream task. Path-JEPA continues to demonstrate strong performance, achieving state-of-the-art results among self-supervised methods while remaining competitive with fully supervised approaches. On NTU60, Path-JEPA achieves 94.4% on X-Sub and 98.2% on X-View, improving over the previous best self-supervised method (GFP) by 0.5% and 1.2% respectively. On the more challenging NTU120 benchmark, Path-JEPA reaches 90.8% on X-Sub and 91.9% on X-Set, representing gains over previous best results. Notably, Path-JEPA outperforms several supervised methods including ST-GCN and 2s-AGCN, demonstrating that signature-based self-supervised pretraining can rival or exceed supervised learning with carefully designed architectures. Compared to recent graph convolutional approaches like CTR-GCN, Path-JEPA achieves comparable performance without requiring explicit graph structure design or message passing operations, suggesting that signature tokenization naturally encodes the relevant spatial relationships through edge and chain tokens.

Table 2: Fine-tuning accuracy (%) on NTU RGB+D datasets. Path-JEPA uses joint positions only. Best results are in **bold**, and second best are underlined.

Method	In.	NTU60		NTU120	
		X-Sub	X-View	X-Sub	X-Set
Supervised methods					
ST-GCN [53]	J	81.5	88.3	70.7	73.2
2s-AGCN [44]	J	88.5	95.1	82.9	84.9
MS-G3D [32]	J	91.5	96.2	86.9	88.4
CTR-GCN [9]	J	92.6	96.9	88.9	90.8
Self-supervised methods					
SkeletonMAE [52]	J	86.6	92.9	76.8	79.1
MotionBERT [58]	J	93.0	97.2	84.8	86.4
MAMP [37]	J	93.1	97.5	90.0	**91.3**
MaskCLR [22]	J	93.5	97.5	90.5	91.9
GFP [48]	J	93.9	97.0	89.6	91.2
S-JEPA [1]	J	93.1	97.6	90.3	91.2
Path-JEPA	J	**94.4**	**98.2**	**90.8**	**91.9**

4.3 Semi-Supervised Learning

To evaluate data efficiency, we perform semi-supervised experiments where only a fraction of labeled data is available for fine-tuning. Table 3 shows results on NTU60 using 1%, 10%, and 60% of training labels. Across all settings, Path-JEPA achieves strong results. For 1% labeled data (approximately 400 samples), Path-JEPA achieves 72.2% on X-Sub and 71.8% on X-View, surpassing S-JEPA by 4.7% and 2.7%, respectively, and matching or exceeding GFP despite using significantly fewer labels. These gains are substantial in the extremely low-data regime, where every percentage point represents successful recognition of dozens of action samples. The improvements remain consistent at 10% labels, with Path-JEPA reaching 89.1% on X-Sub and 92.1% on X-View, maintaining advantages of 0.7% and 0.7% over S-JEPA. At 60% labels, Path-JEPA achieves 92.1% and 95.7%, matching S-JEPA on X-Sub while improving by 0.3% on X-View. These results suggest that signature-based representations provide more transferable features that generalize better with limited labeled data. We hypothesize this stems from the explicit geometric meaning of signature terms: even with few labeled examples, the model can leverage the interpretable displacement and rotation features encoded in signature coefficients to distinguish actions, whereas black-box learned embeddings from S-JEPA or GFP require more labeled data to identify discriminative patterns.

4.4 Transfer Learning

Table 3 evaluates transfer learning where models pretrained on one dataset are fine-tuned on a different target dataset. We use PKU-MMD Phase II as the target and pretrain on NTU60, NTU120, and PKU-MMD Phase I respectively. Path-JEPA achieves the best transfer performance across all source datasets, demonstrating strong generalization of learned representations. When pretraining on NTU60 and transferring to PKU-MMD Phase II, Path-JEPA reaches

Table 3: Semi-supervised learning on NTU60 (accuracy %) with limited labeled data (1/10/60% labels), and transfer learning to PKU-MMD Phase II (PKU-II) where models are pretrained on the source dataset and fine-tuned on PKU-II. Best results are in **bold**, second best are underlined.

Method	Semi-supervised (NTU60)						Transfer to PKU-II		
	X-Sub			X-View			Pretrain → PKU-II		
	1%	10%	60%	1%	10%	60%	NTU60	NTU120	PKU-I
SkeletonMAE [52]	54.4	80.6	85.2	54.6	83.5	88.9	58.4	61.0	62.5
MAMP [37]	66.0	88.0	91.8	68.7	91.5	95.2	70.6	73.2	70.1
S-JEPA [1]	67.5	88.4	**92.1**	69.1	91.4	95.4	71.4	74.2	70.9
GFP [48]	71.8	88.7	–	**72.9**	**92.1**	–	–	–	–
Path-JEPA	**72.2**	**89.1**	**92.1**	71.8	**92.1**	**95.7**	**72.6**	**75.6**	**72.9**

72.6% accuracy, outperforming S-JEPA by 1.2%. Similar improvements of 1.4% and 2.0% are observed when pretraining on NTU120 and PKU-MMD Phase I, respectively. These results indicate that signature-based features capture action characteristics that transfer well across different data distributions, camera setups, and action taxonomies. The mathematical invariances of the signature transform, particularly reparameterization invariance, contribute to this strong transfer capability, as the same action primitives (reaching, grasping, walking) appear across datasets with different recording conditions and temporal sampling rates.

4.5 Ablation Studies

Robustness to Frame Rate Variations To evaluate robustness to frame-rate changes, we apply a controlled resampling protocol on NTU60 X-Sub. For each test sequence, we simulate a target frame rate $f \in \{15, 20, 45, 60\}$ by uniform subsampling (for $f < 30$) or upsampling (for $f > 30$), followed by linear interpolation back to the original sequence length so all methods receive identical input shapes. Table 4 reports accuracy across frame rates (relative to 30fps). Path-JEPA remains stable over 15–60fps, varying by only 1.0% (87.1–88.1). In contrast, S-JEPA drops by 2.9% at 15fps, while SkeletonMAE and MAMP show larger sensitivity to resampling. This behavior aligns with our continuous-time targets: path signatures summarize motion over intervals and satisfy $S(\mathbf{x} \circ \phi) = S(\mathbf{x})$ for monotone reparameterizations ϕ. After lifting sequences to continuous paths via linear interpolation, the resulting signature coefficients are less sensitive to temporal discretization, yielding strong frame-rate robustness without explicit multi-speed augmentation, whereas coordinate- and latent-target baselines must learn such invariances implicitly and degrade more under sampling-rate shifts.

Masking Strategies Table 5 compares different masking strategies. Random masking of 20% of joints achieves 84.1% accuracy. Motion-aware masking im-

Table 4: Robustness to frame-rate variations on NTU60 X-Sub. Accuracy (%) at different frame rates. Best results are in **bold**, and second best are underlined.

Method	15fps	20fps	30fps	45fps	60fps
SkeletonMAE [52]	69.9	73.2	74.8	71.1	70.3
MAMP [37]	81.8	83.5	84.9	82.7	81.2
S-JEPA [1]	82.4	82.9	85.3	84.7	84.5
Path-JEPA	**87.1**	**87.5**	**87.7**	**87.8**	**88.1**

proves it further to 86.8% (+2.7%), confirming that masking informative regions creates more challenging tasks. Our signature-augmented masking without kinematic expansion reaches 86.3%, showing the benefit of leak-free prediction. Finally, kinematic expansion that masks entire subtrees achieves 87.6% (+1.3%), demonstrating that anatomically coherent masking encourages learning of biomechanical constraints. Performance peaks at 20–25% masking, with degradation at higher ratios where insufficient context remains.

Components of Signature Tokenization Table 6 analyzes the contribution of each component in our multi-scale signature tokenization. Using only joint tokens with global signatures achieves 84.4% accuracy on NTU60 X-Sub. Adding local signatures improves performance to 85.8% (+1.4%), demonstrating the value of multi-horizon temporal representation. Including edge tokens further boosts accuracy to 86.4%, showing that pairwise geometric relationships are informative. Finally, adding chain tokens reaches 87.6% (+1.2%), yielding a total 4.3% gain over the baseline.
Using only local signatures without global context performs poorly (83.3%), confirming that long-range temporal dependencies are crucial for distinguishing actions. The hierarchical spatial tokenization progressively incorporates complex geometric relationships, with each level contributing distinct information.

Table 5: Ablation on masking strategies. Linear evaluation on NTU60 X-Sub.

Masking strategy	Acc. (%)
Random (20% joints)	84.1
Motion-aware (S-JEPA)	86.8
Signature-augmented (random seeds)	86.3
+ Kinematic subtree expansion	**87.6**
Kinematic, 10% masking	85.1
Kinematic, 40% masking	84.9

Table 6: Ablation study on tokenization components. Linear evaluation accuracy (%) on NTU60 X-Sub.

Temporal		Spatial hierarchy			Acc. (%)
Global	Local	Joint	Edge	Chain	
✓		✓			84.4
	✓	✓			83.3
✓	✓	✓			85.8
✓	✓	✓	✓		86.4
✓	✓	✓	✓	✓	**87.6**

5 Conclusion

We introduced Path-JEPA, a self-supervised framework that combines path signature representations with masked prediction for skeleton-based action recognition. By constructing multi-scale signature tokens over joint, edge, and chain

paths, Path-JEPA captures displacement, rotation, and higher-order coordination patterns in a geometrically interpretable manner. Our kinematic-aware masking forces the model to reconstruct entire anatomical subtrees from context, yielding temporally robust features that consistently outperform coordinate-based and embedding-based baselines, particularly under frame rate variations and irregular sampling. Despite these benefits, Path-JEPA has limitations. Performance depends on structural hyperparameters such as signature degree and window sizes, which we currently choose heuristically. Future work could co-design the path hierarchy and windowing scheme with the model, allowing adaptive selection of informative paths and scales. Integrating explicit dynamics models or extending Path-JEPA to multi-modal settings (RGB plus skeleton) or other structured time series (biosignals, mesh trajectories) are promising directions. We hope Path-JEPA encourages broader exploration of rough path theory and signature-based representations of continuous time series within modern self-supervised learning frameworks.

References

1. Abdelfattah, M., Alahi, A.: S-jepa: A joint embedding predictive architecture for skeletal action recognition. In: European Conference on Computer Vision. pp. 367–384. Springer (2024)
2. Amraee, S., Singh, A., McCullough, A., Scheithauer, M., Goodwin, M., Ostadabbas, S.: Toward automated behavior understanding in autism: A zero-shot vision-language model approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 8733–8743 (June 2026)
3. Anand, A., Eskandari, A., Rahsno, E., Zulkernine, F.: Asma: Asymmetric spatio-temporal masking for skeleton action representation learning. arXiv preprint arXiv:2602.06251 (2026)
4. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
5. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
6. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
7. Cass, T., Salvi, C.: Lecture notes on rough paths and applications to machine learning. arXiv preprint arXiv:2404.06583 (2024)
8. Chen, L., Nugent, C.D.: Human activity recognition and behaviour analysis. Springer (2019)
9. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)

10. Cheng, J., Shi, D., Li, C., Li, Y., Ni, H., Jin, L., Zhang, X.: Skeleton-based gesture recognition with learnable paths and signature features. IEEE Transactions on Multimedia **26**, 3951–3961 (2024)
11. Degardin, B., Proenca, H.: Human behavior analysis: A survey on action recognition. Applied Sciences **11**(18), 8324 (2021)
12. Do, J., Kim, M.: Skateformer: skeletal-temporal transformer for human action recognition. In: European Conference on Computer Vision. pp. 401–420. Springer (2024)
13. Dong, J., Sun, S., Liu, Z., Chen, S., Liu, B., Wang, X.: Hierarchical contrast for unsupervised skeleton-based action representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 525–533 (2023)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Duan, H., Xu, M., Shuai, B., Modolo, D., Tu, Z., Tighe, J., Bergamo, A.: Skeletr: Towards skeleton-based action recognition in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13634–13644 (2023)
16. Fermanian, A.: Embedding and learning with signatures. Computational Statistics & Data Analysis **157**, 107148 (2021)
17. Friz, P.K., Hairer, M.: A course on rough paths. Springer (2014)
18. Guo, T., Liu, H., Chen, Z., Liu, M., Wang, T., Ding, R.: Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 762–770 (2022)
19. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
20. Jiang, L., Yang, W., Zhang, X., Ni, H.: Gcn-devlstm: Path development for skeleton-based action recognition. arXiv preprint arXiv:2403.15212 (2024)
21. Kidger, P., Bonnier, P., Perez Arribas, I., Salvi, C., Lyons, T.: Deep signature transforms. Advances in neural information processing systems **32** (2019)
22. Lebailly, T., Kips, R., Tinne, T.: Maskclr: Attention-guided contrastive learning for robust action representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
23. LeCun, Y.: A path towards autonomous machine intelligence. vol. 62 (2022)
24. Li, C., Zhang, X., Liao, L., Jin, L., Yang, W.: Skeleton-based gesture recognition using several fully connected layers with path signature features and temporal transformer module. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8585–8593 (2019)
25. Li, L., Wang, M., Ni, B., Wang, H., Yang, J., Zhang, W.: 3d human action representation learning via cross-view consistency pursuit. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4741–4750 (2021)
26. Liao, S., Lyons, T.J., Yang, W., Schlegel, K., Ni, H.: Logsig-rnn: A novel network for robust and efficient skeleton-based action recognition. In: 32nd British Machine Vision Conference. p. 173 (2021)
27. Lin, L., Wu, L., Zhang, J., Liu, J.: Idempotent unsupervised representation learning for skeleton-based action recognition. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)

28. Lin, L., Zhang, J., Liu, J.: Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2363–2372 (2023)
29. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475 (2017)
30. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**(10), 2684–2701 (2019)
31. Liu, Z., Lu, X., Liu, W., Qi, W., Su, H.: Human-robot collaboration through a multi-scale graph convolution neural network with temporal attention. IEEE Robotics and Automation Letters **9**(3), 2248–2255 (2024)
32. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 143–152 (2020)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
34. Lyons, T.: Rough paths, signatures and the modelling of functions on streams. arXiv preprint arXiv:1405.4537 (2014)
35. Lyons, T., McLeod, A.D.: Signature methods in machine learning. arXiv preprint arXiv:2206.14674 (2022)
36. Lyons, T.J., Caruana, M., Lévy, T.: Differential equations driven by rough paths: Ecole d'Eté de Probabilités de Saint-Flour XXXIV-2004. Springer (2007)
37. Mao, Y., Deng, J., Zhou, W., Fang, Y., Ouyang, W., Li, H.: Masked motion predictors are strong 3d action representation learners. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10181–10191 (2023)
38. Mehraban, S., Rajabi, M.J., Iaboni, A., Taati, B.: Stars: Self-supervised tuning for 3d action recognition in skeleton sequences. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2858–2868 (2026)
39. Mo, S., Tong, S.: Connecting joint-embedding predictive architecture with contrastive self-supervised learning. Advances in neural information processing systems **37**, 2348–2377 (2024)
40. Moreno-Pino, F., Arroyo, Á., Waldon, H., Dong, X., Cartea, Á.: Rough transformers: Lightweight and continuous time series modelling through signature patching. Advances in Neural Information Processing Systems **37**, 106264–106294 (2024)
41. Naimi, S., Bouachir, W., Bilodeau, G.A., Mishara, B.: Sstar: Skeleton-based spatio-temporal action recognition for intelligent video surveillance and suicide prevention in metro stations. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 813–823 (2025)
42. Pareek, P., Thakkar, A.: A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. Artificial Intelligence Review **54**(3), 2259–2322 (2021)
43. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
44. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12026–12035 (2019)

45. Su, K., Liu, X., Shlizerman, E.: Predict & cluster: Unsupervised skeleton based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9631–9640 (2020)
46. Sun, H., Chen, Y.: Real-time elderly monitoring for senior safety by lightweight human action recognition. In: 2022 IEEE 16th International Symposium on Medical Information and Communication Technology (ISMICT). pp. 1–6. IEEE (2022)
47. Sun, S., Liu, D., Dong, J., Qu, X., Gao, J., Yang, X., Wang, X., Wang, M.: Unified multi-modal unsupervised representation learning for skeleton-based action understanding. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2973–2984 (2023)
48. Sun, S., Zhang, Z., Dong, J., Cheng, Z., Chang, X., Wang, M.: Towards efficient general feature prediction in masked skeleton modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12212–12221 (2025)
49. Taha, A., Zayed, H.H., Khalifa, M., El-Horbaty, E.S.M.: Skeleton-based human activity recognition for video surveillance. International Journal of Scientific & Engineering Research **6**(1), 993–1004 (2015)
50. Terreran, M., Barcellona, L., Ghidoni, S.: A general skeleton-based action and gesture recognition framework for human–robot collaboration. Robotics and Autonomous Systems **170**, 104523 (2023)
51. Wang, H., Ma, X., Kuang, J., Gui, J.: Heterogeneous skeleton-based action representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19154–19164 (2025)
52. Yan, H., Liu, Y., Wei, Y., Li, Z., Li, G., Lin, L.: Skeletonmae: Graph-based masked autoencoder for skeleton sequence pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5606–5618 (2023)
53. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI Conference on Artificial Intelligence (2018)
54. Yang, W., Lyons, T., Ni, H., Schmid, C., Jin, L.: Developing the path signature methodology and its application to landmark-based human action recognition. In: Stochastic Analysis, Filtering, and Stochastic Optimization: A Commemorative Volume to Honor Mark HA Davis's Contributions, pp. 431–464. Springer (2022)
55. Zhang, H., Wen, M., Wang, L.: Rethinking masked data reconstruction pretraining for strong 3D action representation learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
56. Zheng, L., Shi, L., Zhang, J., Zhang, Y., Liu, J.: Macdiff: Unified skeleton modeling with masked conditional diffusion. In: European Conference on Computer Vision. Springer (2024)
57. Zheng, N., Wen, J., Liu, R., Long, L., Dai, J., Gong, Z.: Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
58. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15929–15939 (2023)