Skeleton-based Action Recognition with Non-linear Dependency Modeling and Hilbert-Schmidt Independence Criterion

Haipeng Chen,[1,2], Yuheng Yang[1,2*], Yingda Lyu[1,3*]

[1]College of Computer Science and Technology, Jilin University,
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University
[3]Public Computer Education and Research Center, Jilin University
chenhp@jlu.edu.cn, yangyh20@mails.jlu.edu.cn, ydlv@jlu.edu.cn,

Abstract

Human skeleton-based action recognition has long been an indispensable aspect of artificial intelligence. Current state-of-the-art methods tend to consider only the dependencies between connected skeletal joints, limiting their ability to capture non-linear dependencies between physically distant joints. Moreover, most existing approaches distinguish action classes by estimating the probability density of motion representations, yet the high-dimensional nature of human motions invokes inherent difficulties in accomplishing such measurements. In this paper, we seek to tackle these challenges from two directions: (1) We propose a novel dependency refinement approach that explicitly models dependencies between any pair of joints, effectively transcending the limitations imposed by joint distance. (2) We further propose a framework that utilizes the Hilbert-Schmidt Independence Criterion to differentiate action classes without being affected by data dimensionality, and mathematically derive learning objectives guaranteeing precise recognition. Empirically, our approach sets the state-of-the-art performance on NTU RGB+D, NTU RGB+D 120, and Northwestern-UCLA datasets.

1 Introduction

Endowing machines with the ability to perceive and recognize human behaviors is very much coveted for various applications ranging from virtual reality to security monitoring (Liu et al. 2022b; Wu et al. 2024b). Consequently, action recognition has attracted much interest, particularly for methods that rely on skeletal data, which is robust against environmental noise and viewpoint changes.

Recent skeleton-based approaches (Yan et al. 2018; Shi et al. 2019a,b) tend to employ Graph Convolutional Networks (GCNs) to model human motion patterns since the hierarchical and tree-like graph structure naturally in the human skeleton. For instance, (Xu et al. 2022) attempts to design sophisticated adjacency matrices, seeking to pursue more nuanced modeling of spatial joint dependencies. (Chi et al. 2022; Zhou et al. 2023) put their efforts into learning discriminative motion features from the skeleton sequence. (Yang et al. 2023) adopts an information-theoretic objective to fully mine task-relevant information while reducing task-irrelevant nuisances.

*Corresponding to Yuheng Yang and Yingda Lyu.

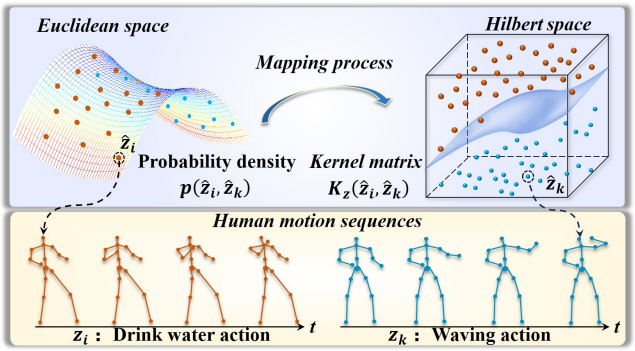

Figure 1: The conceptual diagram illustrates the mapping process of feature representations from Euclidean space into Hilbert space. The mapping process is achieved through the kernel function.

Unfortunately, after a systematic investigation of prior works, we observe that they still suffer from inaccurate action recognition. We conjecture that the reasons are two folds. **(1) First**, current methods (Yan et al. 2018; Li et al. 2019) typically model human motions by performing graph convolutions on pre-defined skeleton graphs. However, these approaches exhibit a severe limitation as they consider only the dependency between physically connected joints while ignoring non-linear dependencies between geometrically separated joints. Although several works (Xu et al. 2022; Song et al. 2022) tend to address this through hierarchical graphs or scaling graphs, they still struggle to effectively model non-linear dependencies between distant joints due to the limited receptive field of graph convolutions. **(2) Second**, given the motion feature representation, existing approaches (Chi et al. 2022; Yang et al. 2023; Zhou et al. 2023; Huang et al. 2023) typically estimate its probability density in conjunction with the category label to identify the action class. A higher probability density indicates a stronger likelihood of the motion feature belonging to the correct class. Nevertheless, a motion sequence is characterized as a time series of skeletal poses, each of which translates to the positions of all joints, resulting in a *high-dimensional* representation. Estimating the probability density of these *high-dimensional* representations in the raw Euclidean space in-

troduces unnecessary training difficulty, which leads to reduced accuracy in action classification.

In this paper, we embrace two key components to tackle these challenges. Technically, **(i)** We propose a dependency refinement method that strips the skeletal structure down to dynamically zoom in on the *joint-to-joint* dependencies of pairwise skeletal joints. Specifically, setting aside the skeletal structure, we explicitly model each pair of joint dependencies with a Gaussian correlation function. By adjusting the kernel width in the Gaussian correlation function, we could fine-grainedly control the influence of distance on joint dependencies. These dependencies are then used to adaptively refine the initial skeletal graph, enabling precise human motion modeling. We also employ an ensemble of networks trained with different kernel widths, seeking to improve the comprehensiveness and accuracy of action recognition. **(ii)** Based on the aforementioned method, we further propose a novel framework that leverages the Hilbert-Schmidt Independence Criterion (HSIC) for facilitating action recognition. First, we utilize a Hilbert kernel function to map high-dimensional motion features from the straightforward Euclidean space to a Hilbert space, as illustrated in Fig. 1. In this space, the HSIC mathematically evaluates the statistical dependence between these features and the corresponding action labels. Second, we theoretically derive the learning objectives, guaranteeing the efficacy of the final action classification. We would like to point out that, since HSIC is defined in terms of kernel method, the process of distinguishing action classes operates in a *dimension-independent* manner.

Thereafter, we conduct extensive experiments on three popular benchmark datasets, namely the NTU RGB+D 60 dataset, the NTU RGB+D 120 dataset, and the Northwestern-UCLA dataset. The empirical results show that our approach consistently and significantly outperforms state-of-the-art performance. To summarize, our key contributions are as follows:

- A dependency refinement method is presented to comprehensively learn the relationships between joints, which simultaneously considers the skeletal connection between adjacent joints and the non-linear dependencies between distant joints.
- We propose a novel action recognition framework, which effectively distinguishes action classes among high-dimensional motion feature representations while deriving learning objectives to ensure the efficacy of the action classification.
- Our method outperforms existing approaches and achieves state-of-the-art performance on three popular benchmark datasets. The implementations have been released, hoping to facilitate future research.

2 Related Work

Skeleton-based action recognition. Previous methods tackle the skeleton-based action recognition task by utilizing Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) (Liu et al. 2016, 2019b), which unfortunately ignore the inherent relationships between joints. Recently, there has been an increasing interest in developing Graph Convolutional Networks (GCNs) (Yan et al. 2018) since the advantages in dealing with irregular graphical structures. FR-Head (Zhou et al. 2023) introduces an auxiliary feature refinement head to acquire discriminative representations of skeletons, aiding in distinguishing ambiguous actions. Stream-GCN (Yang et al. 2023) engages the mutual information loss to maximize the task-relevant information while minimizing the task-irrelevant nuisances for facilitating the action recognition. SkeletonGCL (Huang et al. 2023) introduces graph contrast learning to explore the cross-sequence global context in a fully supervised setting.

Hilbert-Schmidt independence criterion. The Hilbert-Schmidt Independence Criterion (HSIC) is a statistical measure used to assess the independence between two random variables (Gretton et al. 2005). HSIC maps input data into feature vectors and computes their product to evaluate correlation. Its strength lies in uncovering intricate data correlations within a Reproducing Kernel Hilbert Space (RKHS) without being constrained by the dimensionality of the input data (Bertsimas and Koduri 2022). For instance, HSIC-InfoGAN (Liu et al. 2022a) learns unsupervised disentangled representations by directly optimizing the Hilbert-Schmidt Independence Criterion (HSIC) loss, eliminating the need for an auxiliary network. To the best of our knowledge, there is no related research on the utilization of HSIC in the skeleton-based action recognition task.

3 Methodology

Notations. Generally, the human skeleton can be regarded as an interlinked structure comprised of joints and bones. We denote the set of joints as a set of nodes $\mathcal{V} = \{v_1, \ldots, v_N\}$. Additionally, we represent the set of bones as a set of edges $\mathcal{E}$, which can be formulated as an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$. Hence, we conveniently depict a human skeleton as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An action can be represented as a sequence of skeleton poses $\mathcal{X} = \langle x_1, x_2, \cdots, x_T \rangle$, where x_t denotes the skeleton pose at time t and T is the number of frames. Since $x_t \in \mathbb{R}^{N \times C}$, where N is the number of joints, C denotes the dimension of a joint, the 3D motion sequence $\mathcal{X}$ can be formulated as a feature tensor $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$. Presented with a 3D skeleton motion sequence $\mathcal{X}$, we aim to accurately predict its action class label y.

Non-linear Dependency Modeling

Current works (Shi et al. 2019b; Liu et al. 2020; Cheng et al. 2020a) typically learn human motions by modeling the dependencies between adjacent joints. Unfortunately, these methods ignore the relationships between geometrically separated joints. While some approaches introduce adaptive graphs or hierarchical graphs (Lee et al. 2023; Wu et al. 2024a) to mitigate such problems, they are known to have difficulties in capturing the non-linear dependencies between distant joints due to the limited receptive field of graph convolutions. Motivated by these insights, we propose a novel dependency refinement method designed to overcome the limitations imposed by joint distance, allowing for precise human motion modeling. In what follows, we will delve into the specifics of our approach.

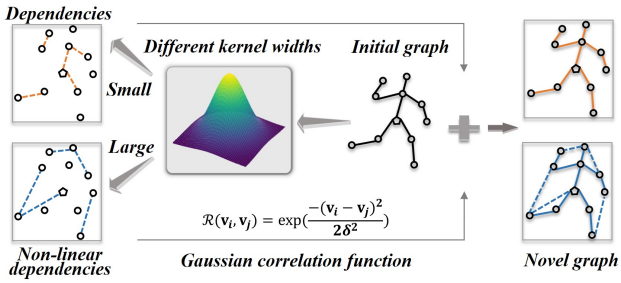

Figure 2: The diagram illustrates the dependency refinement method. Specifically, we utilize the Gaussian correlation function to quantify dependencies between joints and incorporate them into the initial graph. By adjusting the kernel width in the Gaussian function, we could effectively capture the dependencies at both adjacent (orange) and distant (blue) scales.

Technically, for the implementation of graph convolution, we define inputs $\mathcal{X}$ represented by features $\mathbf{X}$ and graph structure $\mathbf{A}$. The update rule of GCNs is given by:

$$\mathbf{X}^{(s)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}\right), \tag{1}$$

where σ indicates an activation function like ReLU, $\mathbf{X}^{(s)} \in \mathbb{R}^{T\times N\times C'}$ denotes the extracted spatial features, $\mathbf{D}$ is the diagonal degree matrix of $\mathbf{A}+\mathbf{I}$, and $\mathbf{W}$ represents the learnable weights for feature transformation. Subsequently, we leverage the Gaussian correlation function $\mathcal{R}(\cdot)$ to model joint dependencies. As shown in Fig. 2, given a pair of joints (v_i, v_j), the features of v_i and v_j can be formulated as $(\mathbf{v}_i, \mathbf{v}_j)$, where $\mathbf{v}_i$, $\mathbf{v}_j \in \mathbb{R}^C$, and i, j indicate the indices of joints. We then input the features $\mathbf{v}_i$ and $\mathbf{v}_j$ into $\mathcal{R}(\cdot)$, formulated as:

$$\mathcal{R}(\mathbf{v}_i, \mathbf{v}_j) = \exp\left(\frac{-(\mathbf{v}_i - \mathbf{v}_j)^2}{2\delta^2}\right), \tag{2}$$

where δ represents the kernel width. Based on the Gaussian correlation function, we obtain the dependencies with a linear transformation ϕ:

$$\mathbf{r}_{ij} = \phi\left(\mathcal{R}(\mathbf{v}_i, \mathbf{v}_j)\right), i, j \in \{1, 2, \cdots, N\}, \tag{3}$$

where $\mathbf{r}_{ij} \in \mathbb{R}^{C'}$ reflects the dependency between v_i and v_j. **Interestingly**, by adjusting the size of δ, we can directly control the degree to which distance affects the dependency. For instance, when performing a complex action such as the "Kicking" action, which requires coordination between the feet, legs, and torso, there are likely to be distant connections between these joints. In such cases, it is necessary to use a large δ so that the dependency decays more slowly as the distance between joints increases. This ensures that the influence of distance on the dependency is weakened, aiming joints that are structurally apart to still exhibit strong dependencies.

Furthermore, we incorporate the dependencies from Equation (3) to refine the initial graph, yielding a novel GCN implementation:

$$\mathbf{X}_c^{(s)} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}(\mathbf{A}_c+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}\right), \tag{4}$$

where $\mathbf{A}_c = \mathbf{A} + \mathbf{W}_c\mathbf{R}$, $\mathbf{W}_c$ denotes the learnable parameters for the dependency, $\mathbf{R} \in \mathbb{R}^{N\times N\times C'}$ denotes the dependency matrix consists of $\mathbf{r}_{ij}$, and $\mathbf{A}_c$ is the novel adjacency matrix obtained in a broadcast manner.

Distinguishing Action Classes For Motion Feature Representations

Many existing methods (Yang et al. 2023; Huang et al. 2023; Zhou et al. 2023) advocate for distinguishing action classes by estimating probability density among multiple motion feature representations. However, factors in the motion sequence, such as *time series*, *3D joint coordinates*, and *the number of joints*, result in a *high-dimensional* feature representation. In such a high-dimensional space, feature representations tend to be sparsely distributed, leaving much of the sampling space unfilled, which makes it difficult to accurately classify the action classes (Majdara and Nooshabadi 2022).

To tackle this problem, we propose a framework based on the Hilbert-Schmidt Independence Criterion. As shown in Fig. 3, our action recognition framework includes a base model for generating motion features and an auxiliary model for producing auxiliary motion information. By incorporating the auxiliary motion information into the motion features, we could explicitly enhance their discriminative ability. These enhanced features are then mapped from Euclidean space to Hilbert space using a Hilbert Reproducing Kernel Function, transforming them into kernel matrices. Finally, we utilize HSIC values derived from these kernel matrices to identify action classes and mathematically formulate training objectives, ensuring effective learning without being hindered by high dimensionality.

Formally, our approach **initiates** by performing joint dependency refinement. **Then**, the base model is trained to encode the motion sequence $\mathcal{X}$ into the motion feature z. **Meanwhile**, the auxiliary model encodes $\mathcal{X}$ into $\tilde{z}$, which is then fed into the classifier to predict motion information $\tilde{y}$. **Thereafter**, we incorporate $\tilde{y}$ into z to obtain the enhanced feature $\hat{z}$ and utilize the Matérn covariance function as the *Hilbert Reproducing Kernel Function* to map $\hat{z}$ into Hilbert Space. The kernel function is:

$$k_\eta(\hat{z}_u, \hat{z}_w) = \alpha\frac{2^{1-\eta}}{\Gamma(\eta)}\left(\frac{\sqrt{2\eta}r}{\ell}\right)^\eta F_\eta\left(\frac{\sqrt{2\eta}r}{\ell}\right), \tag{5}$$

where $\eta > 0$ is the kernel order, F_η is the modified Bessel function of the second kind, $\alpha > 0$ and $\ell > 0$ denote the amplitude and length scale, $\Gamma(\eta)$ is the normalization factor, and $r = \|\hat{z}_u - \hat{z}_w\|_2$. $\{(\hat{z}_u, \hat{z}_w)\}_{u,w=1}^n$ denote u^{th} and w^{th} feature samples. The particular case where $\eta = \frac{3}{2}$ is probably the most commomly used kernel (Williams and Rasmussen 2006):

$$k_{\frac{3}{2}}(\hat{z}_u, \hat{z}_w) = \alpha\left(1 + \frac{\sqrt{3}r}{\ell}\right)\exp\left(-\frac{\sqrt{3}r}{\ell}\right). \tag{6}$$

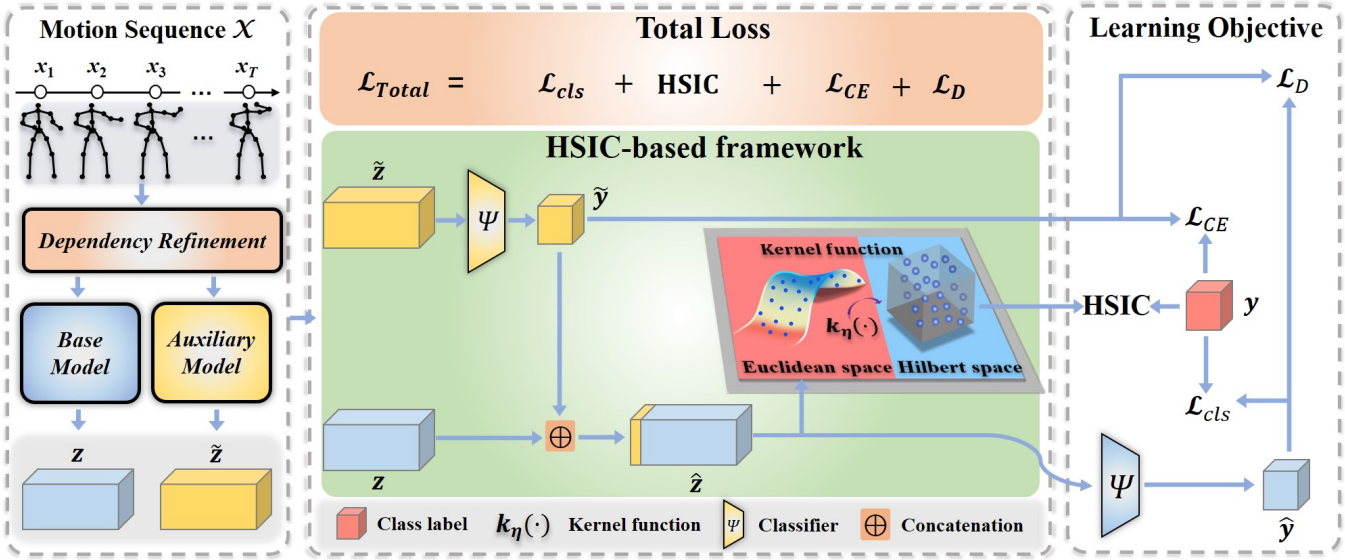

Figure 3: The overall pipeline of our HSIC-based framework aims at recognizing the action classes of the motion sequences. For clarity, we only illustrate the pipeline using a single sequence $\mathcal{S}$ in this figure. The pipeline starts with refining joint dependencies, followed by extracting the motion features z and $\tilde{z}$ from the base model and the auxiliary model, respectively. Subsequently, we feed $\tilde{z}$ to a classifier to obtain the auxiliary information $\tilde{y}$. In order to enhance the discriminative power of z, we incorporate $\tilde{y}$ into z to obtain the augmented feature $\hat{z}$. We then engage a kernel function $k_\eta(\cdot)$ to transform $\hat{z}$ into Hilbert space and derive learning objectives, which effectively avoid the issue arising from the data dimensionality. The entire learning objective $\mathcal{L}_{Total}$ consists of $\mathcal{L}_{cls}$, HSIC, $\mathcal{L}_{CE}$, and $\mathcal{L}_D$.

Subsequently, we could obtain the kernel matrix $K_{\hat{z}}$, which is defined as follows:

$$K_{\hat{z}} = \begin{pmatrix} k_{\frac{3}{2}}(\hat{z}_1, \hat{z}_1) & \dots & k_{\frac{3}{2}}(\hat{z}_1, \hat{z}_n) \\ k_{\frac{3}{2}}(\hat{z}_2, \hat{z}_1) & \dots & k_{\frac{3}{2}}(\hat{z}_2, \hat{z}_n) \\ \vdots & \ddots & \vdots \\ k_{\frac{3}{2}}(\hat{z}_n, \hat{z}_1) & \dots & k_{\frac{3}{2}}(\hat{z}_n, \hat{z}_n) \end{pmatrix}, \tag{7}$$

where each entry $k_{\frac{3}{2}}(\hat{z}_u, \hat{z}_w)$ in Equation (7) denotes the pair-wise kernel value between $\hat{z}_u$ and $\hat{z}_w$. **Finally**, by centering the kernel matrix as described in Equation (7), we calculate the HSIC value, which helps improve the classification efficacy for $\hat{z}$:

$$\text{HSIC}(\hat{z}, y) = \frac{1}{(n-1)^2} \text{tr}(K_{\hat{z}} H K_y H), \tag{8}$$

where $H = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix, $\text{tr}(\cdot)$ denotes the trace operation, y is the class label, and K_y denotes the kernel matrix of y. It is important to note that the HSIC value computation, achieved through the inner product of kernel matrices, is inherently **dimension-agnostic**. Thus we can derive the learning objective combined with Equation (8):

$$\mathcal{L}_H = \mathcal{L}_{cls} + \text{HSIC}(\hat{z}, y), \tag{9}$$

where $\mathcal{L}_H$ is the HSIC-based learning objective and $\mathcal{L}_{cls}$ denotes the empirical classification loss (Chi et al. 2022).

Up to now, we have arrived at the HSIC-based learning objective in Equation (9). Besides this, we further engage in the distillation loss to supervise the alignment between the base model and the auxiliary model. Specifically, to supervise the knowledge distillation from the auxiliary model to the base model, we adopt the distillation method of (Yun et al. 2020). Our distillation loss is defined as:

$$\mathcal{L}_D = \text{KL}(\sigma(f_{\hat{\theta}}(\hat{z}/P)) || \sigma(f_{\tilde{\theta}}(\tilde{z})/P)), \tag{10}$$

where KL is Kullback-Leibler Divergence, σ is the softmax activation, $f(\cdot)$ represents the logit of each model, $\hat{\theta}$ and $\tilde{\theta}$ denote the parameters of the base model and the auxiliary model, and $P > 0$ is the distillation temperature. On the grounds of distillation loss, we incorporate explicit supervision for the auxiliary model using the cross-entropy loss $\mathcal{L}_{CE}$. The entire training objective of our framework is defined as follows:

$$\mathcal{L}_{Total} = \mathcal{L}_H + \mathcal{L}_{CE} + \mathcal{L}_D. \tag{11}$$

Multi-Stream Ensemble

As mentioned earlier, the Gaussian kernel width can regulate the influence of distance on joint dependencies. Therefore, we advocate for training the proposed frameworks with multiple kernel widths to finely capture the non-linear dependencies in different action types. Specifically, a large kernel width is more suitable for modeling actions highlighting distant joint collaborations, such as the "Kicking" action. In contrast, a small kernel width is adaptable for modeling the dependencies between adjacent joints, such as the localized relationship between the head and neck joints during the "Sneeze" action. To this end, we propose feeding the frameworks with joint and bone inputs, each of which is

trained with both small and large kernel widths. During the inference phase, the softmax scores from these frameworks are averaged to obtain final prediction scores. This ensemble method is expected to significantly enhance the comprehensiveness of action recognition

4 Experiments

In this section, we conduct extensive experiments to empirically evaluate the performance of our method on three benchmark action recognition datasets. We aim to answer the following research questions:

- **RQ1:** How is our method comparing to state-of-the-art approaches for skeleton-based action recognition?
- **RQ2:** How much do different components of the proposed method contribute to its performance?
- **RQ3:** What interesting insights and findings can we obtain from the empirical results?

Next, we first present the experimental settings, followed by answering the above research questions one by one.

Experimental Settings

Datasets. We adopt three widely used action recognition benchmark datasets, namely the NTU-RGB+D 60 dataset, the NTU-RGB+D 120 dataset, and the NorthwesternUCLA dataset, to evaluate the proposed method.

NTU-RGB+D (Shahroudy et al. 2016) is designed for the skeleton-based action recognition task. It comprises 56,880 video samples, encompassing 60 action classes performed by 40 volunteers. Each video sample represents a single action and contains a maximum of two subjects. The video sample is recorded using three Microsoft Kinect v2 cameras. The dataset provides two sub-benchmarks: (1) Cross-Subject (X-Sub): data from 20 subjects is used as the training set, while the rest is used as test data. (2) Cross-View (X-View): it divides the training and test sets according to different camera views.

NTU-RGB+D 120 (Liu et al. 2019a) is currently the largest 3D skeleton-based action recognition dataset, which expands on the original NTU-RGB+D dataset by adding 60 additional action classes and 57,600 video samples. It consists of a total of 114,480 samples across 120 action classes, performed by 106 volunteers and captured using three Kinect cameras. The dataset includes two benchmarks: (1) Cross-Subject (X-Sub120) divides the 106 volunteers into two groups, with 53 subjects assigned to the training set and the remaining to the test set. (2) Cross-Setup (X-Set120) are divided into the training and test sets based on their IDs. Samples with even IDs are included in the training set, while samples with odd IDs into the test set.

Northwestern-UCLA (Wang et al. 2014) consists of 1,494 video samples divided into 10 classes. The dataset is recorded using three Kinect cameras and features ten actors. Following the evaluation protocol (Wang et al. 2014), training samples are obtained from the first two cameras, and the remaining camera is used to capture test samples.

Implementation details. We conducted our experiments using two NVIDIA GeForce GTX 3090 GPUs. Our model was implemented using PyTorch 1.11. To train our framework, we employed stochastic gradient descent (SGD) with 0.9 Nesterov momentum. For all datasets, we set the total number of training epochs to 120, with the first 5 epochs dedicated to a warm-up strategy to stabilize the training process. The small Gaussian kernel was set to 1 and the large Gaussian kernel was set to 9. For the NTU-RGB+D and NTU-RGB+D 120 datasets, we set the initial learning rate to 0.1 and applied a decay of 0.1 every 50 epochs. The batch size was set to 128, and the distillation temperature was set to 1.0. For the Northwestern-UCLA dataset, we set the initial learning rate to 0.01, with a decay of 0.1 every 50 epochs. The batch size was set to 32, and the distillation temperature was set to 1.0.

Methods	NTU RGB+D 120	
	X-Sub (%)	X-Set (%)
ST-LSTM (Liu et al. 2016)	55.7	57.9
GCA-LSTM (Liu et al. 2017)	61.2	63.3
ST-GCN (Yan et al. 2018)	70.7	73.2
2s-AGCN (Shi et al. 2019b)	82.9	84.9
DC-GCN+ADG (Cheng et al. 2020a)	86.5	88.1
MS-G3D (Liu et al. 2020)	86.9	88.4
Dynamic GCN (Ye et al. 2020)	87.3	88.6
CTR-GCN (Chen et al. 2021a)	88.9	90.6
MST-GCN (Chen et al. 2021b)	87.5	88.8
Ta-CNN (Xu et al. 2022)	85.4	86.8
EfficientGCN-B4 (Song et al. 2022)	88.3	89.1
STF (Ke, Peng, and Lyu 2022)	88.9	89.9
InfoGCN (Chi et al. 2022)	89.8	91.2
TranSkeleton (Liu et al. 2023)	89.4	90.5
FR-Head (Zhou et al. 2023)	89.5	90.9
Stream-GCN (Yang et al. 2023)	89.7	91.0
SkeletonGCL (Huang et al. 2023)	89.8	91.2
Ours	**90.6**	**91.7**
Ours (Joint)	86.0	87.4
Ours (Bone)	87.6	88.9
Ours (Joint + Bone)	89.3	90.7
Ours (4 ensemble)	**90.6**	**91.7**

Table 1: Comparisons of the Top-1 accuracy (%) with the state-of-the-art methods on the NTU RGB+D 120 dataset.

Comparison with Existing Methods (RQ1)

We first empirically compare the proposed method with the state-of-the-art approaches. The experimental results are summarized in Tables 1–3. Table 1 and Table 2 present results on NTU RGB+D 120 and NTU RGB+D 60 datasets, while Table 3 illustrates results on the Northwestern-UCLA dataset. The proposed method consistently outperforms state-of-the-art approaches on all three datasets. For instance, on the NTU RGB+D 60 dataset, state-of-the-art method (Huang et al. 2023) achieves an accuracy of 92.8%. In comparison, our method achieves a 93.7% accuracy, representing a significant increase of 0.9%. Considering NTU RGB+D 60 is an extensively benchmarked dataset, such improvement is quite hard. The significant improvements validate the effectiveness of our work. Compared to FR-Head with the same ensemble setup (4-ensemble) on the NTU RGB+D 120 dataset, our model outperforms by a margin of 1.1% and 0.8% in cross-subject and cross-set, respec-

Methods	NTU RGB+D 60	
	X-Sub (%)	X-View (%)
IndRNN (Liu et al. 2016)	81.8	88.0
HCN (Liu et al. 2017)	86.5	91.1
ST-GCN (Yan et al. 2018)	81.5	88.3
2s-AGCN (Shi et al. 2019b)	88.5	95.1
SGN (Zhang et al. 2020)	89.0	94.5
AGC-LSTM (Si et al. 2019)	89.2	95.0
DGNN (Shi et al. 2019a)	89.9	96.1
DC-GCN+ADG (Cheng et al. 2020a)	90.8	96.6
Dynamic GCN (Ye et al. 2020)	91.5	96.0
MS-G3D (Liu et al. 2020)	91.5	96.2
MST-GCN (Chen et al. 2021b)	91.5	96.6
CTR-GCN (Chen et al. 2021a)	92.4	96.8
Ta-CNN (Xu et al. 2022)	90.4	94.8
EfficientGCN-B4 (Song et al. 2022)	91.7	95.7
STF (Ke, Peng, and Lyu 2022)	92.5	96.9
InfoGCN (Wang et al. 2022)	93.0	97.1
TranSkeleton (Liu et al. 2023)	92.8	97.0
FR-Head (Zhou et al. 2023)	92.8	96.8
Stream-GCN (Yang et al. 2023)	92.9	96.9
SkeletonGCL (Huang et al. 2023)	92.8	97.1
Ours	**93.7**	**97.3**

Table 2: Comparisons of the Top-1 accuracy (%) with the state-of-the-art methods on the NTU RGB+D 60 dataset.

Methods	Northwestern-UCLA
	Top-1 (%)
Lie Group (Veeriah, Zhuang, and Qi 2015)	74.2
HBRNN-L (Du, Wang, and Wang 2015)	78.5
Ensemble TS-LSTM (Lee et al. 2017)	89.2
SGN (Zhang et al. 2020)	92.5
AGC-LSTM (Si et al. 2019)	93.3
4s Shift-GCN (Cheng et al. 2020b)	94.6
InfoGCN (Chi et al. 2022)	97.0
CTR-GCN (Chen et al. 2021a)	96.5
Ta-CNN (Xu et al. 2022)	96.1
FR-Head (Zhou et al. 2023)	96.8
Stream-GCN (Yang et al. 2023)	96.8
SkeletonGCL (Huang et al. 2023)	96.8
Ours	**97.2**

Table 3: Comparisons of the Top-1 accuracy (%) with the state-of-the-art methods on the NW-UCLA dataset.

tively. These results empirically confirm the superiority of our method in skeleton-based action recognition.

Ablation Study (RQ2)

To investigate the effect of individual components of the proposed method, we examine the classification accuracy of our model with different configurations. All experimental ablation studies are conducted on the X-Sub benchmark of the NTU-RGB+D 120 dataset with joint input modality.

The effects of auxiliary motion information. To study the effects of auxiliary motion information generated from the auxiliary model, we replace the auxiliary model with the different models while maintaining the base model unchanged. Specifically, we use the models proposed in (Yan et al. 2018; Shi et al. 2019b; Cheng et al. 2020a; Chen et al. 2021b; Lee et al. 2023) as the auxiliary models, which exhibit gradually increasing performance. The base model is as proposed in (Yan et al. 2018). The results are summarized in Table 4. From the third line to the seventh line in the table, there is a clear positive correlation between the performance of the auxiliary model and the classification accuracy. This observation indicates that incorporating motion information into motion features plays a beneficial role in enhancing their discriminability, and that better motion information generated from the auxiliary model leads to higher classification accuracy.

Models	# Params.	Acc (%)
Base model (Yan et al. 2018)	1.3M	83.4
Base model + Auxiliary model (Yan et al. 2018)	2.5M	84.4
Base model + Auxiliary model (Shi et al. 2019b)	5.1M	84.5
Base model + Auxiliary model (Cheng et al. 2020a)	4.6M	84.6
Base model + Auxiliary model (Chen et al. 2021b)	4.4M	85.1
Base model + Auxiliary model (Lee et al. 2023)	2.9M	**86.0**

Table 4: Comparisons of classification accuracies when using different auxiliary models to predict auxiliary motion information.

Methods	Acc (%)	Declines (%)
$\mathcal{L}_{Total}$	**86.0**	-
w/o HSIC	85.1	0.9 (↓)
w/o $\mathcal{L}_D$	85.2	0.8 (↓)
w/o HSIC, $\mathcal{L}_D$	84.9	1.1 (↓)

Table 5: Comparison of classification accuracies based on removing the loss term HSIC (w/o HSIC) or removing the loss term $\mathcal{L}_D$ (w/o $\mathcal{L}_D$) from the total loss function $\mathcal{L}_{Total}$.

The effects of learning objectives. We further validate the effect of learning objectives. To confirm that our learning objectives improve test accuracies, we compare the performance of the proposed model trained with different losses by systematically removing each term from the total loss function $\mathcal{L}_{Total}$. The empirical results are presented in Table 5. From the table, we can observe that removing HSIC (w/o HSIC) and removing $\mathcal{L}_D$ (w/o $\mathcal{L}_D$) results in accuracy drops of 0.9% and 0.8%, respectively. Moreover, the performance is significantly decreased by 1.1% when removing both HSIC and $\mathcal{L}_D$ (w/o HSIC, $\mathcal{L}_D$). These findings suggest the effectiveness of the proposed learning objectives in enhancing the entire learning process.

The effects of ensemble streams. To study the effect of ensemble streams, we compare the performance of ensembles of frameworks trained with multiple kernel widths. In Table 1, we observe an improvement in accuracy as the number of ensemble streams increases. On the cross-subject, accuracies of joint+bone and 4 ensemble (ensemble of frameworks trained with both small and large kernel widths using joint and bone input) increased by 3.3% and 4.6%, respectively, compared to the accuracy of using the joint input only. The results imply that multi-stream fusion indeed enhances the comprehensiveness of action recognition, and further increases the recognition performance.

The contributions of each component. We investigate the contribution of each component in the proposed method as illustrated in Table 6. The baseline was constructed by removing the novel graph from the base model in our method

Methods	Acc (%)
Baseline	83.4
+HSIC, $\mathcal{L}_D$	84.7 (1.3↑)
+Novel graph	84.8 (1.4↑)
+HSIC, $\mathcal{L}_D$, Novel graph	**86.0** (2.6↑)

Table 6: Comparisons of classification accuracies when applying each component of our method to the baseline.

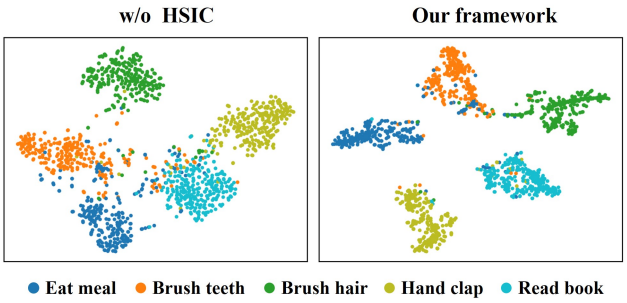

Figure 4: The visualization illustrates feature representations of five randomly selected action classes. We utilize t-SNE for dimension reduction. Each action class is represented by a different color. The five action classes are *Eat meal*, *Brush teeth*, *Brush hair*, *Hand clap*, and *Read book*.

and trained only with $\mathcal{L}_{cls}$. We observe that learning objectives improve baseline accuracy by 1.3%, and the proposed novel graph leads to a 1.4% increase in baseline accuracy. When all proposed components are applied together, the baseline accuracy increases by 2.6%.

Analysis (RQ3)

In this subsection, we present an in-depth visual analysis to study the impact of the HSIC-based learning objective and provide a quantitative analysis of the Gaussian kernel.

Visual analysis of the HSIC-based learning objective. To investigate the effect of the HSIC-based learning objective, we perform a visual analysis of the feature representation $\hat{z}$. We project these feature representations onto a lower dimensional space with t-SNE and compare them using two different methods: the model trained without or with the HSIC-based learning objective.

As shown on the left of Fig. 4, when training without the HSIC-based learning objective, the feature representations exhibit overlapping regions, and each action class lacks clear clustering. In contrast, as depicted on the right of Fig. 4, the proposed method effectively separates the five action classes, with the feature representation of each class becoming more distinct. This indicates that utilizing the proposed HSIC-based learning objective facilitates better distinction between different action classes. We observe similar patterns across all other classes but visualize only five randomly selected action classes for simplicity.

Quantitative analysis of the Gaussian kernel width. To study the effectiveness of the kernel width in the Gaussian

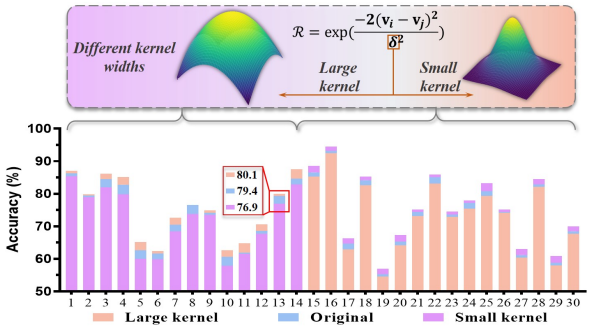

Figure 5: The histograms show the quantitative results of applying the three methods to thirty action classes. The quantitative results of each method are depicted with different colors.

function, we conduct a quantitative analysis of the recognition results. Specifically, we randomly pick thirty action classes and demonstrate their empirical results in Fig. 5. We engage three methods in comparison: the framework training with a large kernel width, the framework with the original graph, and the one with a small kernel width. Taking the "Staggering" action from the figure as an example, we can observe that the framework using a large kernel yields better results compared to that using a small kernel, leading to a 3.2% accuracy increase. This finding suggests that using a large kernel is more suitable for modeling the non-linear dependencies of complex actions that involve multiple distant joint coordination, and vice versa.

5 Limitations and Future Plans

Despite our method achieving state-of-the-art results on all three datasets, there are still certain aspects that warrant further exploration. *Firstly*, expanding our approach to include few-shot learning and unsupervised learning would be fruitful, as it could enhance the applicability of action recognition to real-world scenarios. In our future work, we will concentrate on this aspect. *Secondly*, we plan to further study the application of our feature classification method to other tasks (Tang et al. 2023, 2024; Xu et al. 2024; Liu et al. 2024).

6 Conclusion

In this paper, we have proposed a novel dependency refinement method to model the dependencies between human joints. Concretely, we engage in the Gaussian correlation function to capture non-linear dependencies between any pair of joints, mitigating the influence of joint distance. We further propose a framework to distinguish between high-dimensional motion representations and arrive at well-defined learning objectives that ensure the efficacy of the model. Experimental results demonstrate that our method consistently outperforms state-of-the-art methods on three benchmark datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62276112) and Jilin Province Science and Technology Development Plan Key R&D Project (20230201088GX).

References

Bertsimas, D.; and Koduri, N. 2022. Data-driven optimization: A reproducing kernel Hilbert space approach. *Operations Research*, 70(1): 454–471.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021a. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13359–13368.

Chen, Z.; Li, S.; Yang, B.; Li, Q.; and Liu, H. 2021b. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1113–1122.

Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; and Lu, H. 2020a. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, 536–553. Springer.

Cheng, K.; Zhang, Y.; He, X.; Chen, W.; Cheng, J.; and Lu, H. 2020b. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 183–192.

Chi, H.-g.; Ha, M. H.; Chi, S.; Lee, S. W.; Huang, Q.; and Ramani, K. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20186–20196.

Du, Y.; Wang, W.; and Wang, L. 2015. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1110–1118.

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory: 16th International Conference, ALT 2005, Singapore, October 8-11, 2005. Proceedings 16*, 63–77. Springer.

Huang, X.; Zhou, H.; Feng, B.; Wang, X.; Liu, W.; Wang, J.; Feng, H.; Han, J.; Ding, E.; and Wang, J. 2023. Graph contrastive learning for skeleton-based action recognition. *arXiv preprint arXiv:2301.10900*.

Ke, L.; Peng, K.-C.; and Lyu, S. 2022. Towards to-at spatio-temporal focus for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1131–1139.

Lee, I.; Kim, D.; Kang, S.; and Lee, S. 2017. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, 1012–1020.

Lee, J.; Lee, M.; Lee, D.; and Lee, S. 2023. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10444–10453.

Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; and Tian, Q. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3595–3603.

Liu, H.; Liu, Y.; Chen, Y.; Yuan, C.; Li, B.; and Hu, W. 2023. TranSkeleton: Hierarchical Spatial–Temporal Transformer for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33: 4137–4148.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.-Y.; and Kot, A. C. 2019a. 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.

Liu, J.; Shahroudy, A.; Xu, D.; and Wang, G. 2016. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, 816–833. Springer.

Liu, J.; Wang, G.; Duan, L.-Y.; Abdiyeva, K.; and Kot, A. C. 2017. Skeleton-based human action recognition with global context-aware attention LSTM networks. *IEEE Transactions on Image Processing*, 27(4): 1586–1599.

Liu, X.; Thermos, S.; Sanchez, P.; O'Neil, A. Q.; and Tsaftaris, S. A. 2022a. HSIC-InfoGAN: Learning Unsupervised Disentangled Representations by Maximising Approximated Mutual Information. In *MICCAI Workshop on Medical Applications with Disentanglements*, 15–21. Springer.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

Liu, Z.; Wu, S.; Jin, S.; Liu, Q.; Lu, S.; Zimmermann, R.; and Cheng, L. 2019b. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10004–10012.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022b. Copy Motion From One to Another: Fake Motion Video Generation. *arXiv preprint arXiv:2205.01373*.

Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; and Ouyang, W. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 143–152.

Majdara, A.; and Nooshabadi, S. 2022. Efficient Density Estimation for High-Dimensional Data. *IEEE Access*, 10: 16592–16608.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis.

In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1010–1019.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019a. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7912–7921.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019b. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12026–12035.

Si, C.; Chen, W.; Wang, W.; Wang, L.; and Tan, T. 2019. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1227–1236.

Song, Y.-F.; Zhang, Z.; Shan, C.; and Wang, L. 2022. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tang, F.; Xu, Z.; Huang, Q.; Wang, J.; Hou, X.; Su, J.; and Liu, J. 2023. DuAT: Dual-aggregation transformer network for medical image segmentation. In *PRCV*.

Tang, F.; Xu, Z.; Qu, Z.; Feng, W.; Jiang, X.; and Ge, Z. 2024. Hunting Attributes: Context Prototype-Aware Learning for Weakly Supervised Semantic Segmentation. In *CVPR*.

Veeriah, V.; Zhuang, N.; and Qi, G.-J. 2015. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, 4041–4049.

Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; and Zhu, S.-C. 2014. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2649–2656.

Wang, X.; Dai, Y.; Gao, L.; and Song, J. 2022. Skeleton-based action recognition via adaptive cross-form learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1670–1678.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Transactions on Dependable and Secure Computing*.

Xu, K.; Ye, F.; Zhong, Q.; and Xie, D. 2022. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2866–2874.

Xu, Z.; Tang, F.; Chen, Z.; Zhou, Z.; Wu, W.; Yang, Y.; Liang, Y.; Jiang, J.; Cai, X.; and Su, J. 2024. Polyp-Mamba: Polyp Segmentation with Visual Mamba. In *MICCAI*. Springer.

Yan, S.; Xiong, Y.; Lin, D.; Wang, W.; Wang, L.; and Tan, T. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Yang, Y.; Chen, H.; Liu, Z.; Lyu, Y.; Zhang, B.; Wu, S.; Wang, Z.; and Ren, K. 2023. Action Recognition with Multi-stream Motion Modeling and Mutual Information Maximization. *arXiv preprint arXiv:2306.07576*.

Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; and Tang, H. 2020. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, 55–63.

Yun, S.; Park, J.; Lee, K.; and Shin, J. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13876–13885.

Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; and Zheng, N. 2020. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1112–1121.

Zhou, H.; Liu, Q.; Wang, Y.; Feng, H.; Han, J.; Ding, E.; and Wang, J. 2023. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10608–10617.