InfoGCN: Representation Learning for Human Skeleton-based Action Recognition

Hyung-gun Chi[1]*, Myoung Hoon Ha[2]*, Seunggeun Chi[1], Sang Wan Lee[2], Qixing Huang[3], Karthik Ramani[1,4]
[1]School of Electrical & Computer Engineering, Purdue University, West Lafayette, USA
[2]KAIST, Daejeon, South Korea [3]The University of Texas at Austin, Austin, USA
[4]School of Mechanical Engineering, Purdue University, West Lafayette, USA
{hgchi,sgchi,ramani}@purdue.edu, mh.ha.soar@gmail.com, sangwan@kaist.ac.kr, huangqx@cs.utexas.edu

Abstract

Human skeleton-based action recognition offers a valuable means to understand the intricacies of human behavior because it can handle the complex relationships between physical constraints and intention. Although several studies have focused on encoding a skeleton, less attention has been paid to embed this information into the latent representations of human action. InfoGCN proposes a learning framework for action recognition combining a novel learning objective and an encoding method. First, we design an information bottleneck-based learning objective to guide the model to learn informative but compact latent representations. To provide discriminative information for classifying action, we introduce attention-based graph convolution that captures the context-dependent intrinsic topology of human action. In addition, we present a multi-modal representation of the skeleton using the relative position of joints, designed to provide complementary spatial information for joints. InfoGCN[1] surpasses the known state-of-the-art on multiple skeleton-based action recognition benchmarks with the accuracy of 93.0% on NTU RGB+D 60 cross-subject split, 89.8% on NTU RGB+D 120 cross-subject split, and 97.0% on NW-UCLA.

1. Introduction

Human action recognition is a fundamental problem in computer vision with rich applications, including emergency detection [36], sign language recognition [35], and gesture recognition for VR / AR [57], to name a few. In particular, human action recognition based on the skeleton [6, 7, 19, 44, 58] has attracted much interest in computer vision because of its robustness against a cluttered background. One of the key achievements in skeleton-based action recognition is a graph convolution network (GCN [21]) based approach.

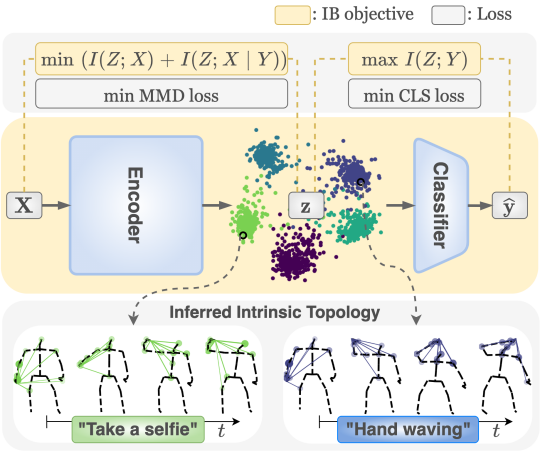

Figure 1. Conceptual diagram of InfoGCN. We propose an IB objective and a corresponding loss to guide our model to learn maximally informative representations for skeleton-based action recognition. The encoder infers an intrinsic topology of joints, which provides contextual information beyond physical connectivity. The colored lines on the bottom indicate inferred intrinsic topology, and the thickness represents the strength of the relation.

This paper introduces a novel skeleton-based prediction framework for action recognition. Our approach advances the state-of-the-art in three critical aspects. The first is the algorithm for representation learning. A large body of works have demonstrated that representation learning considerably influences the performance of machine learning tasks [2, 5, 13, 23, 29, 59, 61]. Our approach is inspired by the information bottleneck (IB) theory [49]. We derive novel IB objective and the corresponding loss to learn the latent representation to be maximally informative to the target variable while compressing the input information conditionally and marginally, as illustrated at the top of Fig. 1. The model learned with the proposed objective performs recognition by encoding implicit and general latent representation, bridging the input-level physical information and action semantics.

The second is the encoding method of the skeleton. Graph representation of the skeleton using bone connectiv-

*These authors contributed equally to this work
[1]Code is available at github.com/stnoah1/infogcn

ity (extrinsic topology) [27, 33, 44, 56, 60] has an inherent limitation: it can ignore possible joint relations, called an intrinsic topology. When we *"take a selfie"*, for instance, there may be an intrinsic relation between the hand holding a phone and the upper body since we jointly move them to locate the upper body on the screen of the phone (as the inferred intrinsic topology from our model in Fig. 1). The intrinsic topology of joints [40] provides contextual information to recognize human action. In this context, we develop a novel self-attention based graph convolution (SA-GC) module, to extract the intrinsic graph structure when encoding a sequence of the skeleton. As shown at the bottom of Fig. 1, for the similar poses that appear in different actions, the inferred topologies can be different based on their behavioral contexts.

Lastly, we propose a multi-modal skeleton representation by utilizing the joints' relative positions. It provides complementary spatial information of a joint. An ensemble of the models trained with the representations drastically improves recognition performance.

By coupling the aforementioned three proposals, we introduce a new learning framework for skeleton-based action recognition named InfoGCN. To verify the effectiveness of our approach, we perform empirical evaluations in skeleton-based action recognition and compare our results with competitive baselines on three popular benchmark datasets: NTU RGB+D 60 & 120 [30, 42], and NW-UCLA [55]. Experimental results show that our model achieves state-of-the-art performance on all three datasets in terms of accuracy. Analysis shows that the learned latent representation of action adheres to the proposed IB constraints, and the context-dependent intrinsic topology is inferred adaptively depending on the behavioral context.

Our contributions are as follows:

- **Information Bottleneck Objectives**. We introduce a novel learning objective based on IB that aims to learn an efficiently compressed latent representation of an action.
- **Self-Attention based Graph Convolution**. We propose an SA-GC module that infers a context-dependent intrinsic topology in spatial modeling of a skeleton.
- **Multi-Modal Representation**. We present a multi-modal representation of a skeleton for the model ensemble that drastically improves action recognition performance.
- **Empirical Verification**. Extensive experiments demonstrate the advantages of our work. InfoGCN achieves state-of-the-art performance on the three datasets in skeleton-based action recognition.

2. Related Works

In the early stages of deep learning-based approaches for skeleton-based action recognition, convolutional neural networks (CNNs) [10, 32, 46] and recurrent neural networks (RNNs) [11, 25, 31, 53] were the standard models to adopt. However, the capability of these methods were limited as they did not explicitly exploit the structural topology of the joints. Since the introduction of GCN [21], various approaches exploiting the graph structure of extrinsic topology have been introduced [26, 33, 43, 56]. Various graphs, including a spatio-temporal graph [56] and a directed graph [43], have been proposed to model the skeleton. Multi-scale graph convolutions [26, 33] have been presented to capture long-range dependencies of joints. Nonetheless, these methods are not able to represent the intrinsic topology, limiting the ability to capture the contextual information of the action.

Recent works [27, 44] focus on joint topology modeling that can infer intrinsic relations. AS-GCN [27] and 2s-AGCN [44] propose methods that adaptively learn joint relations from the data. However, since the captured topologies are independent of a pose, it has difficulties encoding the context of an action whose pose changes over time. CTR-GCN [6] is similar to our work in terms of context-dependent intrinsic topology modeling. In contrast to our work, CTR-GCN focuses on embedding joint topology in different embedding channels. Meanwhile, unlike previous studies that focus only on spatio-temporal feature aggregation of the skeleton, to the best of our knowledge, InfoGCN is the first approach that leverages an information-theoretic objective to better represent latent information.

3. InfoGCN

InfoGCN is a novel learning framework that predicts the action class for a given sequence of skeletons. In this section, we first derive an IB-based learning objective and the corresponding loss (Sec. 3.1). In addition, we introduce a neural architecture (Sec. 3.2) and multi-modal representations of the skeleton for the model ensemble (Sec. 3.3). The overall learning scheme is presented at the end. Note that all notations used in this section are summarized in the Appendix.

3.1. Information-Bottleneck Objectives

The goals of this section are to define an objective based on IB for learning a latent representation from a sequence of skeletons and to derive its variational bound and tractable loss. The proposed formulation can be applied to other problems such as human motion prediction and self-supervised learning.

3.1.1 Learning Objective

We aim to design a stochastic latent variable Z containing compressed information with respect to the input variable X (a sequence of skeletons), while preserving maximum information for the target variable Y (an action label). This

constrained optimization can be transformed to an unconstrained one with a Lagrangian multiplier: $\max_Z I(Z;Y) - \beta_1 I(Z;X)$, where $I(\cdot;\cdot)$ is mutual information and β_1 is the Lagrangian multiplier. As in prior works [2, 13], we assume that the relation of variables follows the graphical model $Z \leftarrow X \leftrightarrow Y$, and the only accessible content is the stochastic encoder $p(z|x)$. In infoGCN we propose the following objective equivalent to maximizing the prior IB objective (See Appendix):

$$R(Z) = I(Z;Y) - \lambda_1 I(Z;X) - \lambda_2 I(Z;X|Y), \quad (1)$$

where λ_1 and λ_2 are control parameters. The first term $I(Z;Y)$ forces Z to be informative enough for predicting Y. The second term ensures that Z is concise. The third term allows the latent variable Z to be compressed with respect to the input variable X when a class is given. Our objective adopts the combination of the compression regularizer terms from VIB [2] and CEB [13] while retaining the IB philosophy. Our derived objectives are more general than those of [2, 13] while incorporating the previous objectives as special cases (VIB when $\lambda_1 = 0$ and CEB when $\lambda_2 = 0$).

3.1.2 Variational Bound

Here we derive the variational bound of our IB objective (Eq. (1)). The variational bound of each term of $R(Z)$ is derived following recent studies [2, 4, 37], which estimate mutual information using tractable variational bound and deep learning techniques. We obtain the variational lower bound for $I(Z;Y)$ using a variational classifier $q(y|z)$:

$$I(Z;Y) \geq E_{p(x,y)p(z|x)}[\log q(y|z)] + H(Y), \quad (2)$$

where the first term of RHS corresponds to log-likelihood and the second term of RHS is constant when the underlying data generating distribution is fixed, so it does not affect the optimization. Following [13, 17], we define $r(z)$ as the variational marginal and $r(z|y)$ as the variational class conditional marginal. We obtain the variational upper bounds of $I(Z;X)$ and $I(Z;X|Y)$ as in [13, 17]

$$I(Z;X) \leq E_{p(x)p(z|x)}[\log \frac{p(z|x)}{r(z)}], \quad (3)$$

$$I(Z;X|Y) \leq E_{p(x)p(z|x)p(y|x)}[\log \frac{p(z|x)}{r(z|y)}]. \quad (4)$$

Substituting Eqs. (2) to (4) into Eq. (1), we have

$$R(Z) \geq E_{p(x,y)p(z|x)}[\log q(y|z)] - \lambda_1 E_{p(x)p(z|x)}[\log \frac{p(z|x)}{r(z)}] - \lambda_2 E_{p(x)p(z|x)p(y|x)}[\log \frac{p(z|x)}{r(z|y)}]. \quad (5)$$

Derivations of Eqs. (2) to (4) are provided in the Appendix.

3.1.3 Training Loss

We define the loss function for training InfoGCN from the lower bound of our objective function (Eq. (5)). The first term of Eq. (5) can be approximated by the empirical loss of the prediction network combining the encoder and the classifier:

$$\begin{aligned} \mathcal{L}_{\text{CLS}} &= -E_{p(x,y)p(z|x)}[\log q(y|z)] \\ &\approx -\frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} E_{p(z|x_i)}[\log q(y_i|z)], \end{aligned} \quad (6)$$

where $\mathcal{D} = \{(x_i, y_i)\}$ is a given dataset.

The second term of Eq. (5) can be further decomposed following [16, 34].

$$E_{p(x)p(z|x)}[\log \frac{p(z|x)}{r(z)}] = I(Z;X) + D_{\text{KL}}(p(z)||r(z)) \quad (7)$$

We perform two simplifications. The first is to drop $I(Z;X)$ to prioritize that Z contains compressed information with respect to X [19, 34]. The second is to replace the intractable KL-diverse term $D_{\text{KL}}(p(z)||r(z))$ with the tractable Maximum-Mean Discrepancy (MMD [12, 14, 28]), which has proven to be valid and effective in the literature [61]. We set the domain and codomain as a Euclidean space and feature map as an identity. This gives us the following marginal-MMD loss:

$$\begin{aligned} \mathcal{L}_{\text{mMMD}} &= D_{\text{MMD}}(p(z)||r(z)) \\ &= ||\mu_{p(z)} - \mu_{r(z)}||_2^2, \end{aligned} \quad (8)$$

where $\mu_{p(z)} = \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in D} E_{p(z|x_i)}[z]$; $\mu_{r(z)}$ is the mean of the variational marginal distribution $r(z)$.

The last term of bound in Eq. (5) is decomposed following the same procedure with Eq. (7), and we have the following conditional-marginal-MMD loss as

$$\begin{aligned} \mathcal{L}_{\text{cmMMD}} &= D_{\text{MMD}}(p(z|y)||r(z|y)) \\ &= \sum_y ||\mu_{p(z|y)} - \mu_{r(z|y)}||_2^2, \end{aligned} \quad (9)$$

where $\mu_{p(z|y)} = \frac{1}{|\mathcal{D}_y|} \sum_{x_i, y_i \in D_y} E_{p(z|x_i)}[z]$, and $\mathcal{D}_y = \{x_i, y_i | y_i = y\}$.

Finally, we have a total loss function to train our model:

$$\mathcal{L}_{\text{TOTAL}} = \mathcal{L}_{\text{CLS}} + \lambda_1 \mathcal{L}_{\text{mMMD}} + \lambda_2 \mathcal{L}_{\text{cmMMD}}. \quad (10)$$

3.2. Neural Architecture

We introduce a neural architecture that can model the context-dependent intrinsic topology of joints by leveraging the self-attention mechanism [51]. It includes an encoder-classifier structure as illustrated in Fig. 2.

3.2.1 The Importance of Learning Intrinsic Topology

We present the importance of intrinsic topology by showing that GC using only extrinsic topology can cause severe inefficiency and information loss in terms of message passing. Assume that both hand joints have an intrinsic relation due to bilateral symmetricity of the body structure. These two leaf nodes must pass messages through the physically connected path to transmit information to each other. When transferring information following the mechanism of GC, it requires an increase in the depth of the network in proportion to the length of the shortest path for message passing, which implies serious inefficiency in information exchange between two nodes.

Furthermore, information loss can happen. GC performs a nonlinear transformation after averaging features of neighbor nodes. If the feature vectors representing the information for the nodes are not linearly independent, it is not easy to reconstruct the information of each neighbor after averaging. Let α be the minimum portion of diluted information of a node caused by vector composition. If the distance between two nodes has an intrinsic relation l, information can be transmitted with the maximum ratio of $(1-\alpha)^l$. When $\alpha > 0$, the longer l, the more information can be diluted.

A straightforward approach is to increase the convolution kernel size by powering the adjacency matrix as in [19, 26], but this is not ideal because it cannot dynamically model possible intrinsic topologies. A better solution is to adaptively infer the joint relations required to change information. Therefore, we propose an architecture that utilizes a self-attention mechanism to capture the intrinsic topology.

3.2.2 Architecture Overview

The encoder is composed of an embedding block and a stack of $L = 9$ encoding blocks followed by a global average pooling layer. The embedding block transforms a sequence of the skeleton to initial joint representations. Then, encoding blocks extract spatio-temporal features from the initial joint representations. We leverage the method of the reparameterization trick in VAE [20]. With an auxiliary independent random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{z}$ is sampled as $\mathbf{z} = \mu + \mathbf{\Sigma}\epsilon$, where the mean μ and diagonal covariance matrix $\mathbf{\Sigma}$ of the multivariate Gaussian distribution are inferred from the output of the encoder. The trick makes the model trainable by estimating unbiased gradients in an end-to-end fashion with gradient-based optimization.

A classifier, composed of a single linear layer and softmax function, converts the latent vector $\mathbf{z}$ to the model parameter of the categorical distribution.

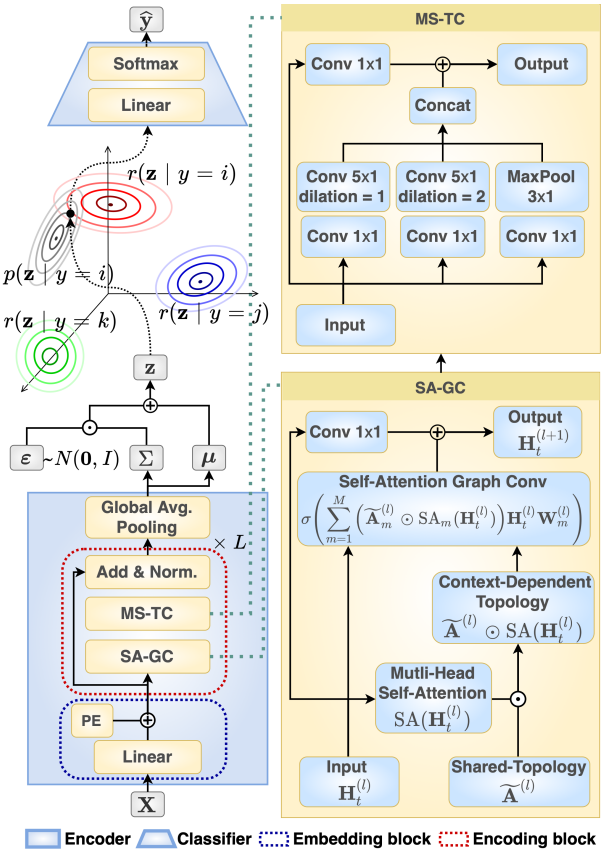

Figure 2. InfoGCN Architecture. We guide our neural architecture to learn the class conditional representation of skeleton-based action with the information bottleneck objective. The model is composed of an encoder and a classifier. The encoder with the SA-GC module captures context-dependent intrinsic joint topology to better represent action.

3.2.3 Embedding Block

The human skeleton can be represented as a graph $\mathcal{G}(V, E)$ with joints as a set of N vertices V and bone as edges E. Edges can be represented as an adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$, where $\mathbf{A}_{i,j} = 1$ if joints i and j are physically connected, otherwise 0. A sequence of skeleton graphs is represented as a joint feature tensor $\mathbf{X} \in \mathbb{R}^{T \times N \times C}$, where T is the number of total frames of the skeleton and C is the feature dimension.

The embedding block linearly transforms the joint features to $D^{(0)}$ dimensional vectors with learnable parameters and then adds positional embeddings (PE) to inject positional information of joints. We adapt learnable PE, which is shared across times.

$$\mathbf{H}_t^{(0)} = \texttt{Linear}(\mathbf{X}_t) + PE, \tag{11}$$

where $\mathbf{H}_t^{(0)}, PE \in \mathbb{R}^{N \times D^{(0)}}$; t is the time index.

3.2.4 Encoding Block

The core of our encoding block consists of two sub-modules: a Self-Attention based Graph Convolution (SA-GC) module for spatial modeling and a Multi-Scale Temporal Convolution (MS-TC) module for temporal modeling. The input and hidden representation of joints are encoded sequentially with an SA-GC, an MS-TC, a residual connection, and a layer normalization [3] (See Fig. 2).

Spatial Modeling. We propose a novel module SA-GC to infer context-dependent intrinsic topology. Before describing SA-GC, we revisit vanilla GC [21], which is composed of two processes; 1) average neighborhood vertex features and 2) linearly transform aggregated features. The update rule of hidden representation for GC is as follows

$$\mathbf{H}_t^{(l+1)} = \sigma(\hat{\mathbf{A}}\mathbf{H}_t^{(l)}\mathbf{W}^{(l)}), \tag{12}$$

where normalized adjacency matrix $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{D}^{-\frac{1}{2}}$, $\mathbf{D}$ is the diagonal degree matrix of $\mathbf{A}+\mathbf{I}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{D^{(l)}\times D^{(l+1)}}$ is learnable parameters of the l-th layer, and $\sigma(\cdot)$ indicates nonlinear activation function like ReLU [1].

SA-GC utilizes the self-attention [51] of joint features to infer intrinsic topology and uses the topology as a neighborhood vertex information for the GC. A self-attention is an attention mechanism that relates different joints of the body. Considering all possible joint relations, SA-GC infers positive and bounded weight, called self-attention map, to represent the strength of the relation. We linearly project joint representation $\mathbf{H}_t$ to queries and keys of D' dimensions with learned matrices $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{D\times D'}$ to get a self-attention map.

$$\text{SA}(\mathbf{H}_t) = \texttt{softmax}(\frac{\mathbf{H}_t\mathbf{W}_K(\mathbf{H}_t\mathbf{W}_Q)^T}{\sqrt{D'}}) \tag{13}$$

In addition to the self-attention map, we let SA-GC learn a topology $\tilde{\mathbf{A}}$ shared over time and instance as in [6, 44]. The shared-topology and self-attention map have M multi-head to make the model jointly attend from different representation subspaces. For a head in $1 \le m \le M$, we combine the shared topology $\tilde{\mathbf{A}}_m \in \mathbb{R}^{N\times N}$ with the self-attention map $\text{SA}_m(\mathbf{H}_t) \in \mathbb{R}^{T\times N\times N}$ to obtain the intrinsic topology.

$$\tilde{\mathbf{A}}_m \odot \text{SA}_m(\mathbf{H}_t) \in \mathbb{R}^{T\times N\times N}, \tag{14}$$

where $\odot$ indicates broadcasted element-wise product. We employ $D' = D/8$ and $M = 3$ in this work.

SA-GC utilizes $\tilde{\mathbf{A}}_m \odot \text{SA}_m(\mathbf{H}_t)$ as neighborhood information for GC. The overall update rule of joint representation is formulated as

$$\mathbf{H}_t^{(l+1)} = \sigma\left(\sum_{m=1}^{M}(\tilde{\mathbf{A}}_m^{(l)} \odot \text{SA}_m(\mathbf{H}_t^{(l)}))\mathbf{H}_t^{(l)}\mathbf{W}_m^{(l)}\right) \tag{15}$$

We employ a residual connection [15] with 1×1 convolution around the SA-GC module.

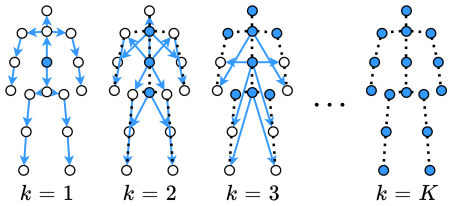

Figure 3. Illustration of multi-modal representation of the skeleton. Arrows describe the k-th mode representation of pointed vertices. As in [44], we define the joint close to the center of mass as the source joint, and the joint far from it as the target joint. Blue dots denote vertices with no corresponding source.

Temporal Modeling. To model the temporal feature of the human skeleton, we adopt the MS-TC module [6, 33] as shown in Fig. 2. This module consists of three convolution branches with different combinations of kernel sizes and dilation rates. The outputs of convolution branches are concatenated. A residual connection with 1×1 convolution is around this module.

3.3. Ensemble with Multi-Modal Representation

In this section, we introduce a generalized form of well-known skeleton representation such as *bone* and *joint*, which we call multi-modal representation. We train our model with each modal representation and ensemble upon inference. The representation provides complementary features using the relative position of joints. See Fig. 3 for illustration.

Shi et al. [44] introduce *bone information*, which is defined as a vector pointing toward its target joint from its source joint that are physically connected, as shown at $k = 1$ in Fig. 3. Previous works [6, 33, 44] show that the ensemble of models trained with *bone* and *joint information* drastically improves action recognition performance, implying that these different representations of skeleton are complementary. We propose multi-modal skeleton representation to define additional representations, based on the fact that *bone information* is a linear transformation of *joint*. In detail, we generalize joint-bone relation at time t as

$$\tilde{\mathbf{X}}_t^{(k)} = (\mathbf{I} - \mathbf{P}^k)\mathbf{X}_t, \tag{16}$$

where $\mathbf{P} \in \mathbb{R}^{N\times N}$ denotes a binary matrix that contains source-target relations of the skeleton graph, $\mathbf{P}_{ij} = 1$ if the i-th joint is the source of the j-th joint, otherwise 0. We set the row corresponding to the center of mass in P as a zero vector so that it does not have a source joint. We refer to $\tilde{\mathbf{X}}_t^{(k)}$ as the k-th mode representation of a skeleton. The representations with different k values provide distinct spatial features for a joint. We define $K = \max_v d(v) + 1$ for $v \in V$, where $d(v)$ gives the shortest distance in the number of hops between the vertex v and the center of mass. Then, if $k = 1$, the k-th mode representation $\tilde{\mathbf{X}}_t^{(k)}$ corresponds

Methods	Acc (%)	
	X-Sub	X-View
ST-GCN [56]	81.5	88.0
AS-GCN [27]	86.8	94.2
2s-AGCN [44]	88.5	95.1
SGN [60]	89.0	94.5
DGNN [43]	89.9	96.1
ST-TR-agcn [39]	90.3	96.3
Shift-GCN [9]	90.7	96.5
DC-GCN+ADG [8]	90.8	96.6
PA-ResGCN-B19 [47]	90.9	96.0
DDGCN [22]	91.1	**97.1**
Dynamic GCN [58]	91.5	96.0
MS-G3D [33]	91.5	96.2
MST-GCN [7]	91.5	96.6
CTR-GCN [6]	92.4	96.8
Ours	**93.0**	**97.1**

Methods	Acc (%)	
	X-Sub	X-Set
SGN [60]	79.2	81.5
2S-AGCN [44]	82.9	84.9
ST-TR-agcn [39]	85.1	87.1
Shift-GCN [9]	85.9	87.6
DC-GCN+ADG [8]	86.5	88.1
MS-G3D [33]	86.9	88.4
PA-ResGCN-B19 [47]	87.3	88.3
Dynamic GCN [58]	87.3	88.6
MST-GCN [7]	87.5	88.8
CTR-GCN [6]	88.9	90.6
Ours (Joint)	85.1	86.3
Ours (Bone)	87.3	88.5
Ours (Joint + Bone)	88.5	89.7
Ours (4 ensemble)	89.4	90.7
Ours (6 ensemble)	**89.8**	**91.2**

Methods	Acc (%)
Lie Group [52]	74.2
Actionlet ensemble [54]	76.0
HBRNN-L [11]	78.5
Ensemble TS-LSTM [24]	89.2
AGC-LSTM [45]	93.3
Shift-GCN [9]	94.6
DC-GCN+ADG [8]	95.3
CTR-GCN [6]	96.5
Ours	**97.0**

Table 1. Comparative results on NTU RGB+D 60 [42] (*left*), NTU RGB+D 120 [30] (*middle*), and NW-UCLA [55] (*right*). We evaluate our model in terms of classification accuracy (%). The performance of baseline methods is based on their papers. Bold figures indicate the best value for each dataset. X-Sub, X-view, and X-Set represent cross-subject, cross-view, and cross-setup splits, respectively.

to the bone as defined in [44] and if $k = K$, joint since $\mathbf{P}^K = \mathbf{0}$. For instance, at $k = 1$ in Fig. 3, a joint of the center of mass is represented as a blue dot, so K is equal to 5 in this case.

3.4. Learning Framework

This section describes the overall training regime of InfoGCN. Sequences of the skeletons are batched together after being resized to 64 frames as in [6]. The model is updated to minimize the total loss (Eq. (10)) using SGD optimizer with a momentum coefficient 0.9. We set the $\mu_{r(z)}$ to be $\mathbf{0}$ so that $\mathcal{L}_{\text{mMMD}}$ behaves as a regularizer of the norm of $\hat{\mu}_{p(z)}$. We set the $\mu_{r(z|y)}$ of each action class as random orthogonal vectors [41] with a scale of 3. During the training, we estimate $\hat{\mu}_{p(z)}$ and $\hat{\mu}_{p(z|y)}$ by averaging marginal and class conditional marginal latent vectors of a mini-batch, respectively. Also, we employ label smoothing [48] of value 0.1. During inference, we ensemble models that trained with different k-mode representations as the multi-stream ensemble in [6, 33, 44].

4. Experiments

To demonstrate the advantages of InfoGCN, we conduct skeleton-based action recognition on three large-scale datasets. We compare our model with strong baselines and conduct ablation studies to examine the effect of individual components. Our model is implemented with PyTorch [38], and trained and tested using an NVIDIA RTX A6000 GPU. Further details of our experimental setups are described in the Appendix.

4.1. Datasets

NTU RGB+D. NTU RGB+D 60 [42] is a large-scale 3D human activity dataset having 56,880 videos composed of 60 action classes. NTU RGB+D 120 [30] is an extended version of NTU RGB+D 60 with an additional 60 extra action classes and contains 114,480 videos. As recommended by [30, 42], we report classification accuracies under cross-subject and cross-view settings for NTU RGB+D 60 and cross-subject and cross-setup settings for NTU RGB+D120.

NW-ULCA. NW-ULCA [55] has 1,494 videos of 10 different actions simultaneously captured from three cameras. We use data from the first two cameras for training and the other for testing as in [55].

4.2. Experimental Results

We compare our results with previous state-of-the-art methods in Table 1. We set K to be 8 for NTU RGB+D 60 & 120 and 6 for NW-UCLA. In the middle of Table 1, ensemble of pose and motion with multi-modal representation $k = \{1, K\}$ and $\{1, 2, K\}$ are denoted as 4 and 6 ensemble, respectively. Here, motion means joint movement between two subsequent time frames. In the left and right of Table 1, we report the results of 6 ensemble. On all three datasets, InfoGCN achieves state-of-the-art performance, validating the effectiveness of our work. With the same ensemble setup with CTR-GCN (4 ensemble) [6] on NTU-RGB+D 120, our model outperforms CTR-GCN by a margin of 0.5% and 0.1% in cross-subject and cross-set, respectively (see the middle of Table 1). These results empirically verify the advantage of InfoGCN in skeleton-based action recognition.

4.3. Ablation Studies

To analyze the effect of individual components of InfoGCN, we examine the classification accuracy of different configurations of our model. All experimental ablation studies are conducted on NTU RGB+D 120 cross-subject split with joint information ($k = K$).

Methods	Acc (%)
$\mathcal{L}_{\text{TOTAL}}$	**85.1**
w/o $\mathcal{L}_{\text{mMMD}}$	84.6
w/o $\mathcal{L}_{\text{cmMMD}}$	84.6
w/o $\mathcal{L}_{\text{mMMD}}$, $\mathcal{L}_{\text{cmMMD}}$	84.3

Methods	Acc (%)
$\hat{\mathbf{A}} \odot \text{SA}(\mathbf{H}_t)$	**85.1**
$\text{SA}(\mathbf{H}_t)$	84.7
$\tilde{\mathbf{A}}$	84.5
$\hat{\mathbf{A}}$	82.8

Table 2. Comparison of classification accuracies based on (*left*) removing $\mathcal{L}_{\text{mMMD}}$ or $\mathcal{L}_{\text{cmMMD}}$ from the total loss and (*right*) different topology inference methods.

Methods	Acc (%)	
	4-Stream	Multi-modal
Baseline	88.8	89.2 (0.4↑)
+ $\mathcal{L}_{\text{mMMD}}$, $\mathcal{L}_{\text{cmMMD}}$	89.1 (0.3↑)	89.4 (0.6↑)
+ SA-GC	89.1 (0.3↑)	89.5 (0.7↑)
+ $\mathcal{L}_{\text{mMMD}}$, $\mathcal{L}_{\text{cmMMD}}$, SA-GC	89.4 (0.6↑)	**89.8** (1.0↑)

Table 3. Comparisons of classification accuracies when applying the proposed components of InfoGCN to the baseline.

MMD loss. We first validate the effect of MMD losses derived from IB objective in Sec. 3.1.3. To confirm that our objective increases the effect of generalization directly leading to improved test accuracies, we compare the performance of our model trained with different losses by removing each of the loss terms from $\mathcal{L}_{\text{TOTAL}}$ as shown in the left of Table 2. We observe that the performance of our model trained without both $\mathcal{L}_{\text{cmMD}}$ and $\mathcal{L}_{\text{mMMD}}$ drops 0.8% compared with the original one. The performance of InfoGCN trained without $\mathcal{L}_{\text{cmMMD}}$ and without $\mathcal{L}_{\text{mMMD}}$ drop both 0.5%, confirming that MMD losses guide our model to learn better representation for action classification.

Context-dependent topology. We compare the classification accuracy of models using different topology inference methods as shown in the right of Table 2. We use each inferred topology as neighborhood information for GC. We see that the models with adaptive topology inference methods such as $\tilde{\mathbf{A}}$, $\text{SA}(\mathbf{H}_t)$, and $\hat{\mathbf{A}} \odot \text{SA}(\mathbf{H}_t)$ outperform static extrinsic topology $\hat{\mathbf{A}}$. Moreover, context-dependent topology $\hat{\mathbf{A}} \odot \text{SA}(\mathbf{H}_t)$ is superior to other methods, demonstrating the effectiveness of SA-GC.

Multi-modal representation. We compare the performance of ensembles of models trained with different combination of multi-modal representations. In the middle of Table 1, we observe the improvement of performance as the number of modalities for ensemble increases. On the cross-subject, accuracies of joint+bone, 4, and 6 ensembles are improved by 3.4%, 4.3%, and 4.7%, respectively, compared to the accuracy of joint only. This implies that multi-modal representations increase the diversity of input features and the number of corresponding trained models, further maximizing the effect of the ensemble. The accuracies tend to be saturated after 6 mode ensemble (See Appendix). As k increases, the number of vertices without a source increase, which are marked as blue dots in Fig. 3, and they do not provide distinctive features.

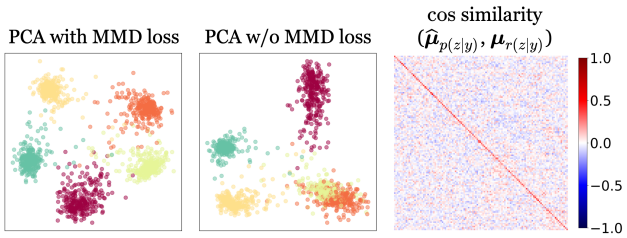

Figure 4. (*Left-middle*) PCA projection of latent representation to 2D when trained with or without MMD loss. We randomly select five action classes for visualization from NTU RGB+D 120 dataset. Different colors indicate different classes. (*Right*) cosine similarity between $\hat{\mu}_{p(z|y)}$ and $\mu_{r(z|y)}$. Each row and column indicates different classes.

Contribution of each Component We scrutinize the contribution of each InfoGCN component as shown in Table 3. The baseline was built by replacing the SA-GC with the GC in [56] from our model (Sec. 3.2) and trained only with $\mathcal{L}_{\text{CLS}}$. The 4-stream ensemble [6, 7, 9, 43] is adopted for the baseline to compare with the ensemble results of models trained with multi-modal representations as inputs. We observe that SA-GC and MMD losses ($\mathcal{L}_{\text{mMMD}}$, $\mathcal{L}_{\text{cmMMD}}$) both improve baseline accuracy by 0.3% each. In addition, when we adopt multi-modal representations for the model ensemble, the accuracy improved by 0.4% compared to the 4-stream ensemble baseline.

5. Analysis

We conduct an in-depth analysis of the proposed learning objective and context-dependent intrinsic topology. All analyses are based on the model trained with the joint ($k = K$) on NTU RGB+D 120 cross-subject split.

5.1. Information Bottleneck Constraint

To validate the effect of the proposed objective, we trained our model with or without MMD loss and compared action representations by principal component analysis (PCA [50]) as shown in Fig. 4. Latent representation learned with MMD loss presents a denser and non-overlapped class conditional distribution that seems more discriminative on the subspace spanned by the first two principal components than those without MMD loss. We observe similar patterns in all other classes but visualize only five categories for simplicity.

We compare the cosine similarity between the $\hat{\mu}_{p(z|y)}$ averaged on the test set and the $\mu_{r(z|y)}$ as shown in the right of Fig. 4. We see that the diagonal elements of the matrix have values closer to 1 while off-diagonal entries are near 0, indicating $\hat{\mu}_{p(z|y)}$ and $\mu_{r(z|y)}$ are well aligned as intended. We attribute the performance gain to the observation that MMD loss successfully constrains the mean of class conditional representation to be close to $\mu_{r(z|y)}$, which is set to be class-wise orthogonal.

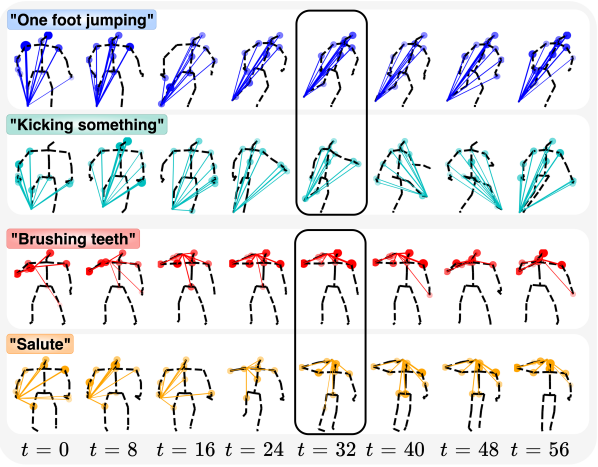

Figure 5. Examples of context-dependent intrinsic topology of SA-GC. Colored lines indicate inferred topology from a specified joint (a hand or a foot) to all the other joints. The thickness of the colored lines and the size of circles on joints are proportional to the strength of the inferred relation. Black bounding boxes indicate similar poses with different intrinsic topologies.

5.2. Context-Dependent Intrinsic Topology

Behavioral Context. Fig. 5 gives examples of the topology inferred by SA-GC. We observed that similar poses (grouped by bounding boxes) could have different intrinsic topologies depending on their behavioral context, which may be the reason why our model better distinguishes behavior patterns. For instance, *One foot jumping* and *kicking something* at $t = 32$ in Fig. 5 have similar poses. Their intrinsic topologies (colored lines), however, are distinctive. Attention from the joint of the right foot to the left arm is stronger in action *kicking* than *jumping*. One possible explanation is that the left hand moves in the opposite direction to the right foot to balance the body when kicking something, so they are strongly coupled. Whereas the joints of the right foot and the left hand are not much related when jumping with one foot. We also observed the marginal attention to see effectiveness of intrinsic topology for describing context as shown in the left of Fig. 6. The joint of the right hand in action *taking a selfie* has a large magnitude of attention, while the right foot is strongly attended in action *one foot jumping*. This observation is intuitive since the right hand is actively involved in action *taking a selfie*, and the foot is mostly used in action *one foot jumping*. Moreover, multi-head attentions provide different behavioral contexts for action as shown in the right of Fig. 6.

Asymmetric Message Passing. Unlike the extrinsic topology, self-attention maps are inferred to be asymmetric as shown in Fig. 6. Since the amount of message passing between the joints can differ depending on the direction, SA-GC can transfer information between joints efficiently, overcoming the limitation of GC described in Sec. 3.2.1.

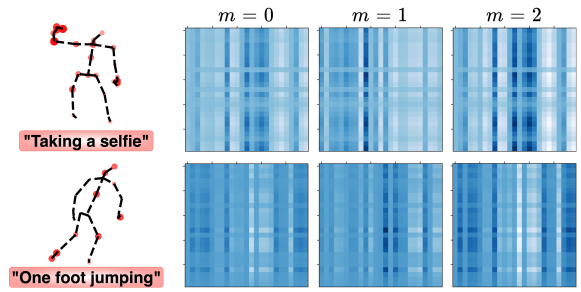

Figure 6. Examples of self-attention maps in Eq. (13) along with the magnitude of attention for each skeleton joint. The size of circles in the skeleton joint represents the magnitude of attention, which is defined as the sum of the column of each joint in the self-attention map. We visualize self-attention maps with different heads m of the last encoding block. The darker the color, the higher the value is in the self-attention map.

6. Limitations

Despite the state-of-the-art performance of InfoGCN on three datasets, its effectiveness in the dataset with a large number of classes (i.e., 400 actions for kinetics-400 [18]) remains to be tested. This warrants further investigation to demonstrate the model's capacity to deal with the larger number of classes and larger batch size. Besides, it would be interesting to extend our approach to meta-learning and self-supervised learning to exploit unlabeled data. Lastly, the application of InfoGCN is confined to human skeleton modeling. That being said, we should note that InfoGCN can be applied to any structured data, such as the motion of particles and articulated objects.

7. Conclusion

We present an information bottleneck-based representation learning framework, InfoGCN, for skeleton-based human action recognition. It is built based on a variational bound of the information-theoretic objective, encouraging the mean of the class conditional marginal to be nearly orthogonal. We propose a novel self-attention-based graph convolution module, SA-GC, and demonstrate that it can effectively glean behavioral context information from data using the inferred intrinsic topology. We further introduce a multi-modal representation of the human skeleton for model ensemble. Notably, our framework achieves state-of-the-art performance on three popular benchmark datasets for skeleton-based action recognition.

Acknowledgement This work was partially supported by US National Science Foundation (FW-HTF 1839971), the National Research Foundation (NRF-2019M3E5D2A01066267), KAIST-KT Joint Research Center (Genie Brain: Developing an abstraction and reasoning engine thinking like a human brain), and Institute for Information communications Technology Planning & Evaluation (IITP, No.2019-0-01371, Development of brain-inspired AI with human-like intelligence). We also acknowledge the Feddersen Chair Funds for Professor Karthik Ramani.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 5

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 1, 3

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540. PMLR, 2018. 3

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 1

[6] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 1, 2, 5, 6, 7

[7] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1113–1122, 2021. 1, 6, 7

[8] Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 536–553. Springer, 2020. 6

[9] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 6, 7

[10] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015. 2

[11] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015. 2, 6

[12] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015. 3

[13] Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020. 1, 3

[14] Arthur Gretton, Karsten Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander J Smola. A kernel method for the two-sample problem. *arXiv preprint arXiv:0805.2368*, 2008. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, 2016. 3

[17] HyeongJoo Hwang, Geon-Hyeong Kim, Seunghoon Hong, and Kee-Eung Kim. Variational interaction information maximization for cross-domain disentanglement. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 8

[19] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 1, 3, 4

[20] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[21] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1, 2, 5

[22] Matthew Korban and Xin Li. Ddgcn: A dynamic directed graph convolutional network for action recognition. In *European Conference on Computer Vision*, pages 761–776. Springer, 2020. 6

[23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse Image-to-Image Translation via Disentangled Representations. *arXiv:1808.00948 [cs]*, Aug. 2018. arXiv: 1808.00948. 1

[24] Inwoong Lee, Doyoung Kim, Seoungyoon Kang, and Sanghoon Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1012–1020, 2017. 6

[25] Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. Rnn fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*, pages 833–850. Springer, 2016. 2

[26] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. Spatio-temporal graph convolution for skeleton based action recognition. *CoRR*, abs/1802.09834, 2018. 2, 4

[27] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019. 2, 6

[28] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015. 3

[29] Alexander H. Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation. *arXiv:1809.01361 [cs]*, Oct. 2018. arXiv: 1809.01361. 1

[30] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 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. 2, 6

[31] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1656, 2017. 2

[32] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 2

[33] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 143–152, 2020. 2, 5, 6

[34] Alireza Makhzani and Brendan Frey. Pixelgan autoencoders. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1972–1982, 2017. 3

[35] Ashish S Nikam and Aarti G Ambekar. Sign language recognition using image based hand gesture recognition techniques. In *2016 online international conference on green engineering and technologies (IC-GET)*, pages 1–5. IEEE, 2016. 1

[36] Cosmas Ifeanyi Nwakanma, Fabliha Bushra Islam, Mareska Pratiwi Maharani, Dong-Seong Kim, and Jae-Min Lee. Iot-based vibration sensor data collection and emergency detection classification using long short term memory (lstm). In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 273–278. IEEE, 2021. 1

[37] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[38] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[39] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding*, 208:103219, 2021. 6

[40] Sujan Sarker, Sejuti Rahman, Tonmoy Hossain, Syeda Faiza Ahmed, Lafifa Jamal, and Md Atiqur Rahman Ahad. Skeleton-based activity recognition: Preprocessing and approaches. *Contactless Human Activity Analysis*, 200:43, 2021. 2

[41] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. 6

[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2, 6

[43] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7912–7921, 2019. 2, 6, 7

[44] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 1, 2, 5, 6

[45] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. 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*, pages 1227–1236, 2019. 6

[46] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[47] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1625–1633, 2020. 6

[48] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 6

[49] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 1

[50] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 7

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 5

[52] Vivek Veeriah, Naifan Zhuang, and Guo-Jun Qi. Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 4041–4049, 2015. 6

[53] Hongsong Wang and Liang Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 499–508, 2017. 2

[54] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Learning actionlet ensemble for 3d human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(5):914–927, 2013. 6

[55] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 2, 6

[56] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action

recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2, 6, 7

[57] LI Yang, Jin Huang, TIAN Feng, WANG Hong-An, and DAI Guo-Zhong. Gesture interaction in virtual reality. *Virtual Reality & Intelligent Hardware*, 1(1):84–112, 2019. 1

[58] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 55–63, 2020. 1, 6

[59] Xiaoming Yu, Yuanqi Chen, Thomas Li, Shan Liu, and Ge Li. Multi-mapping Image-to-Image Translation via Learning Disentanglement. *arXiv:1909.07877 [cs]*, Dec. 2019. arXiv: 1909.07877. 1

[60] Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1112–1121, 2020. 2, 6

[61] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017. 1, 3