Appendices of
InfoGCN: Representation Learning for Human Skeleton-based Action Recognition

A. Derivation from Objective to Loss

IB Objective The objective $I(Z;Y) - \beta_1 I(Z;X)$ based on the information bottleneck (IB) can be transformed into $I(Z;Y) - \lambda_1 I(Z;X) - \lambda_2 I(Z;X|Y)$ using a relation obtained from either the chain rule for conditional mutual information or interaction information.

Conditional mutual information is the mutual information of two random variables given a conditional variable. The chain rule for (conditional) mutual information can be described as

$$I(X;Y|Z) = I(X;Y,Z) - I(X;Z) \tag{1}$$
$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z). \tag{2}$$

On the other hand, interaction information is a generalized version of mutual information for multiple variables. Using the relationship between mutual information and conditional mutual information, interaction information for three variables X, Y, and Z can be defined as follows

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z) \tag{3}$$
$$= I(X;Z) - I(X;Z|Y) \tag{4}$$
$$= I(Y;Z) - I(Y;Z|X). \tag{5}$$

Since interaction information has a symmetric property for variables as in mutual information, it does not matter which variable is condition.

Our model follows the graphical assumption ($Z \leftarrow X \leftrightarrow Y$) in prior studies [1,5] based on the information bottleneck objective; thus, the variables Z and Y become independent when X is observed. i.e, $Z \perp Y|X$. From this, we can derive the following relation from the chain rule or interaction information:

$$I(Z;X) = I(Z;Y) + I(Z;X|Y). \tag{6}$$
$$I(Z;Y) = I(Z;X) - I(Z;X|Y). \tag{7}$$

In detail, using the chain rule and conditional independent assumption, we have

$$I(Z;X|Y) = I(Z;X,Y) - I(Z;Y) \tag{8}$$
$$\Rightarrow I(Z;X|Y) = I(Z;X,\cancel{Y}) - I(Z;Y) \tag{9}$$
$$\Rightarrow I(Z;X) = I(Z;Y) + I(Z;X|Y). \tag{10}$$

Meanwhile, using conditional independent assumption and the equality between Eq. (4) and Eq. (5), we can derive

$$I(Z;X) - I(Z;X|Y) = I(Z;Y) - I(Z;Y|X) \tag{11}$$
$$\Rightarrow I(Z;X) - I(Z;X|Y) = I(Z;Y) - \cancel{I(Z;Y|X)} \tag{12}$$
$$\Rightarrow I(Z;X) = I(Z;Y) + I(Z;X|Y). \tag{13}$$

From the IB objective, the equivalent objective that we propose can be derived in two ways using one of the relations in Eqs. (6) and (7). Here, we introduce the derivation of our objective using Eq. (6). First, we decompose the mutual information term $I(Z;X)$ in the IB objective with a new coefficient β_2, replacing one of the decomposed terms using the relation in Eq. (6), and expand the objective as follows

$$I(Z;Y) - \beta_1 I(Z;X) \tag{14}$$
$$= I(Z;Y) - \beta_1(1 - \beta_2 + \beta_2) I(Z;X) \tag{15}$$
$$= I(Z;Y) - \beta_1\beta_2 I(Z;X) - \beta_1(1-\beta_2)\big(I(Z;Y) + I(Z;X|Y)\big) \tag{16}$$
$$= \big(1 - \beta_1(1-\beta_2)\big) I(Z;Y) - \beta_1\beta_2 I(Z;X) - \beta_1(1-\beta_2) I(Z;X|Y). \tag{17}$$

Since a scaling of the objective is independent of optimization, we can normalize the coefficients as

$$I(Z;Y) - \frac{\beta_1\beta_2}{1-\beta_1(1-\beta_2)} I(Z;X) - \frac{\beta_1(1-\beta_2)}{1-\beta_1(1-\beta_2)} I(Z;X|Y). \tag{18}$$

In particular, if $\beta_2 = 1$, the second term in Eq. (18) becomes zero, which reduces Eq. (18) to the IB objective in VIB [1], and if $\beta_2 = 0$, the third term becomes zero, which reduces Eq. (18) to the objective in CEB [5]. We simplify it by introducing new coefficients as

$$I(Z;Y) - \lambda_1 I(Z;X) - \lambda_2 I(Z;X|Y), \tag{19}$$

where coefficient $\lambda_1 = \frac{\beta_1\beta_2}{1-\beta_1(1-\beta_2)}$ and $\lambda_2 = \frac{\beta_1(1-\beta_2)}{1-\beta_1(1-\beta_2)}$.

Varitional Bound for IB Objective Let us consider mutual information $I(Z;Y)$, which is defined as

$$I(Z;Y) = E_{p(z,y)}\big[\log \frac{p(z,y)}{p(z)p(y)}\big] \tag{20}$$

$$= E_{p(z,y)}\big[\log \frac{p(y|z)}{p(y)}\big]. \tag{21}$$

We assume that our model follows the graphical model, and the only accessible content is $p(z|x)$ as in [1, 5]. In mutual information, the decoding distribution $p(y|z)$ entails the expectation over x, derived from the accessible content $p(z|x)$, which is generally intractable, since it is difficult to directly access the underlying distribution. We resolve the intractability of $p(y|z)$ by introducing a variational approximation, which is denoted as $q(y|z)$.

$$I(Z;Y) \geq E_{p(z,y)}[\log \frac{q(y|z)}{p(y)}] \tag{22}$$

$$= E_{p(z,y)}[\log q(y|z)] + H(Y). \tag{23}$$

The inequality holds since $D_{KL}\big(p(y|z)||q(y|z)\big) \geq 0$. The entropy term $H(Y)$ can be ignored because it is irrelevant to optimization when the dataset is given.

Similarly, the definition of mutual information $I(Z;X)$ is as follows

$$I(Z;X) = E_{p(z,x)}[\log \frac{p(z,x)}{p(z)p(x)}] \tag{24}$$

$$= E_{p(z,x)}[\log \frac{p(z|x)}{p(z)}]. \tag{25}$$

By marginalizing out x from the joint distribution $p(z,x)$, $p(z)$ can be obtained. However, since it is generally intractable, a variational approximation to $p(z)$, called $r(z)$, is used to approximate $p(z)$.

$$I(Z;X) \leq E_{p(z,y)}[\log \frac{p(z|x)}{r(z)}]. \tag{26}$$

The inequality holds since $D_{KL}\big(p(z)||r(z)\big) \geq 0$.

Mutual information $I(Z;X|Y)$ is a conditional version of $I(Z;X)$. The definition of $I(Z;X|Y)$ is in Eq. (27), and its variational bound is derived by introducing a variational distribution $r(z|y)$.

$$I(Z;X|Y) = E_{p(z,x,y)}[\log \frac{p(z|x,y)}{p(z|y)}] \tag{27}$$

$$\leq E_{p(z,x,y)}[\log \frac{p(z|x)}{r(z|y)}]. \tag{28}$$

Decomposition We decompose the variational bound on $I(Z;X)$ as in [8, 13].

$$E_{p(x)p(z|x)}[\log \frac{p(z|x)}{r(z)}]$$

$$= E_{p(x)p(z|x)}[\log \frac{p(z|x)}{p(z)} + \frac{p(z)}{r(z)}] \tag{29}$$

$$= E_{p(x)p(z|x)}[\log \frac{p(z|x)}{p(z)}] + E_{p(z)}[\log \frac{p(z)}{r(z)}] \tag{30}$$

$$= I(Z;X) + KL(p(z)||r(z)). \tag{31}$$

Similarly, the variational bound on $I(Z;X|Y)$ can be decomposed as follows

$$E_{p(x,y)p(z|x)}[\log \frac{p(z|x)}{r(z|y)}]$$

$$= E_{p(x,y)p(z|x)}[\log \frac{p(z|x,y)}{p(z|y)} + \frac{p(z|y)}{r(z|y)}] \tag{32}$$

$$= E_{p(x,y)p(z|x)}[\log \frac{p(z|x,y)}{p(z|y)}] + E_{p(x,y)p(z|x)}[\log \frac{p(z|y)}{r(z|y)}] \tag{33}$$

$$= I(Z;X|Y) + KL(p(z|y)||r(z|y)). \tag{34}$$

Maximum Mean Discrepancy Regularizer Maximum Mean Discrepancy (MMD) [4, 6, 10] is a measure of the divergence of two distributions. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a positive semi-definite kernel with a feature map $\phi : \mathcal{X} \mapsto \mathcal{H}$, where $\mathcal{X}$ is domain and $\mathcal{H}$ is the corresponding reproducing kernel Hilbert space (RKHS). The MMD of distributions p and q are:

$$D_{\text{MMD}}(p(x)||q(x)) = ||E_p(x)[\phi(x)] - E_q(x)[\phi(x)]||^2_{\mathcal{H}}. \tag{35}$$

It is 0 when $p = q$ and greater than 0 when $p \neq q$. If we set $\mathcal{X} = \mathcal{H} = \mathbb{R}^d$ and $\phi(x) = x$, Eq. (35) is reduced to $||E_p(x)[x] - E_q(x)[x]||^2$, which is the regularizer function in InfoGCN.

The regularizer $\mathcal{L}_{\text{mMMD}}$ minimizes the divergence of the distributions $p(z)$ and $r(x)$ using MMD. It does not regularize over the variance or shape of the distribution, but the position of the mean to be adjusted to origin. From the perspective of deep learning, regularizing the distribution center to the origin alleviates the factors that increase the size of the bias parameters of the network. In addition, the regularizer $\mathcal{L}_{\text{cmMMD}}$ minimizes the divergence of the distributions $p(x|z)$ and $r(x|z)$ using MMD. Similar to $\mathcal{L}_{\text{mMMD}}$, no regularization is performed on the variance or shape of the conditional marginal distribution, while only the positions of the means are regularized to be orthogonal between the classes. The MMD losses ($\mathcal{L}_{\text{mMMD}}$, $\mathcal{L}_{\text{cmMMD}}$) are expected to have a generalization effect on classification through regularization along with classification loss that is directly learned by maximizing the likelihood of $q(y|z)$.

	Notations	Descriptions
Random variables	X	Input variable (sequence of skeleton)
	Z	Latent variable
	Y	Target variable (action label)
Control parameters	β_1	Lagrangian multiplier for IB objective in [1]
	λ_1 and λ_2	Control parameter of our objective
Variational distributions	$R(Z)$	Our objective
	$q(y\|z)$	Variational classifier
	$r(z)$	Variational marginal
	$r(z\|y)$	Variational class conditional marginal
Datasets	$\mathcal{D}$	Dataset
	$\mathcal{D}_y$	class conditional Dataset
Losses	$\mathcal{L}_{\text{CLS}}$	Classificiation loss
	$\mathcal{L}_{\text{mMMD}}$	Marginal-MMD loss
	$\mathcal{L}_{\text{cmMMD}}$	Conditional-marginal-MMD loss
	$\mathcal{L}_{\text{TOTAL}}$	Total loss
Graph	V	Vertices of skeleton graph
	E	Edges of skeleton graph
Constants	C	Skeleton feature dimension
	T	The number of total frame
	N	The number of skeleton joint
Variables and learnable parameters	$\mathbf{X}$	Joint feature
	$\mathbf{A}$	Adjacency matrix
	$\hat{\mathbf{A}}$	Normalized adjacency matrix
	$\tilde{\mathbf{A}}$	(Learnable) shared topology
	$\mathbf{P}$	Skeleton bone source-target relation matrix
	$\mathbf{D}$	Diagonal degree matrix of $\mathbf{A}+\mathbf{I}$
	$\mathbf{H}$	Joint representation
	$\mathbf{W}$	Learnable weight of SA-GC
Reparameterization parameters	$\mathbf{z}$	Latent vector of action
	μ	Mean of multi-variate Gaussian Distribution
	Σ	Diagonal covariance matrix of multi-variate Gaussian Distribution
	ϵ	Auxilary independent random noise
Hyper-parameters	K	The maximum number of hops from skeleton center of mass.
	M	The number of head in MSA
	D	Dimension of embedding block
	D'	Dimension per head

Table 1. Summary of notations

B. Notations of InfoGCN

In Table 1, we present all notations that used to describe InfoGCN in Sec 3.

C. Implementation Details

Code We provide the implementation code for InfoGCN in a publically accessible repository `github.com/stnoah1/infogcn` to reproduce the experimental results. The code includes instructions for data acquisition, preprocessing, dependencies, and exact commands for the experiments.

Model Architecture Table 3 shows the output size of each block of InfoGCN. The number of joints N is 25 for NTU RGB+D 60 [11] & 120 [11] and 20 for NW-UCLA [16]. The number of action classes is 60 for NTU RGB+D 60 & 120 and 10 for NW-UCLA. To infer the latent vectors, we adopt three fully connected (`FC`) layers. The first `FC` layer transforms pooled aggregated features from encoding blocks to the vector of latent dimension. Then, the `MeanFC` and `CovFC` layers output the mean and covariance of the multivariate Gaussian distribution, which models conditional latent distribution.

Datasets	K	# Subject	# Class	# Joint (N)	Train / Test Split	# Train	# Test	# Total
Northwestern-UCLA [16]	6	10	10	20	-	1,020	474	1,494
NTU RGB+D 60 [14]	8	40	60	25	Cross-Subject	40,091	16,487	56,578
	8	40	60	25	Cross-View	37,646	18,932	56,578
NTU RGB+D 120 [11]	8	106	120	25	Cross-Subject	63,026	50,919	113,945
	8	106	120	25	Cross-Setup	54,468	59,447	113,945

Table 2. Datasets summary.

Datasets In Table 2, we present metadata of datasets that are used for the experiments, including the number of data points and action classes of each dataset. We also provide URLs and terms and conditions for the usage of datasets[1]. One can obtain the datasets: NTU RGB+D 60 & 120 [2] and NW-UCLA [3]. The terms and conditions of NTU RGB+D 60 & 120 provided by the authors are as follows

> *"Both 'NTU RGB+D' and 'NTU RGB+D 120' are released for academic research only, and are free to researchers from educational, or research institutes for non-commercial purposes. The use of these two datasets is governed by the following terms and conditions: Without the expressed permission of the ROSE Lab, any of the following will be considered illegal: redistribution, derivation or generation of a new dataset from this dataset, and commercial usage of any of these datasets in any way or form, either partially or in its entirety. For the sake of privacy, images of all subjects in any of these datasets are only allowed for demonstration in academic publications and presentations. All users of 'NTU RGB+D' and 'NTU RGB+D 120' action recognition datasets agree to indemnify, defend and hold harmless, the ROSE Lab and its officers, employees, and agents, individually and collectively, from any and all losses, expenses, and damages."*

Preprocessing Protocol For NTU RGB+D 60 & 120 datasets, we adopt the preprocessing procedure of [17] and align the spine of the skeletons based on the view-invariant transformation [15], so that skeletons are perpendicular to the ground. For the NW-UCLA dataset, we follow the methods presented in [2, 3].

Training We train the InfoGCN with 110 epochs and use warm-up [7] for the early 5 epochs as in [2]. The learning rate is initialized to 0.1 for NTU RGB+D 60 & 120 and 0.05 for NW-UCLA and decayed with a factor of 0.1 at 90 and 100 epochs. For NTU RGB+D 60 & 120 datasets, we use the weight decay of 5×10^{-4} and the loss coefficients of $\lambda_1 = 1 \times 10^{-4}$ and $\lambda_2 = 1 \times 10^{-1}$. For the NW-UCLA dataset, weight decay of 4×10^{-4} and the loss coefficients of $\lambda_1 = 1 \times 10^{-1}$ and $\lambda_2 = 1 \times 10^{-1}$ are used. The batch size is set up to be about twice the number of classes so that the mini-batch can contain two data on average for each class; 32 for NW-UCLA, 128 for NTU RGB+D 60, and 256 for NTU RGB+D 120.

[1] We failed to get a license information of NW-UCLA online.

[2] https://rose1.ntu.edu.sg/dataset/actionRecognition/

[3] https://wangjiangb.github.io/my_data.html

	Blocks	Output Size
Input		$64 \times N \times 3$
Encoder	Embedding Block	$64 \times N \times 64$
	Encoding Block 1	$64 \times N \times 64$
	Encoding Block 2	$64 \times N \times 64$
	Encoding Block 3	$64 \times N \times 64$
	Encoding Block 4	$32 \times N \times 128$
	Encoding Block 5	$32 \times N \times 128$
	Encoding Block 6	$32 \times N \times 128$
	Encoding Block 7	$16 \times N \times 256$
	Encoding Block 8	$16 \times N \times 256$
	Encoding Block 9	$16 \times N \times 256$
	Global Average Pooling	$1 \times 1 \times 256$
	`FC`	256
	`MeanFC`	256
	`CovFC`	256
	Latent Vector	256
Classifier		# Action Class

Table 3. Shape of output tensor for each block of InfoGCN. The output size of encoding and embedding blocks denote the number of time frames $\times$ the number of joints $\times$ the embedding dimension.

D. Additional Experimental Results

D.1. Performance Plot of Previous SoTAs

Fig. 1 plots action classification accuracies of the previous methods and ours on three datasets. We note that the performance gain of InfoGCN is considerable. (Performance gains of ours versus average of the previous state-of-the-art records on NTU120 X-sub: 0.9% vs 1.0% and X-view: 0.6% vs 1.13% / NTU60 X-sub: 0.6% vs 0.5% / NW-UCLA: 0.5% vs 1.6%).

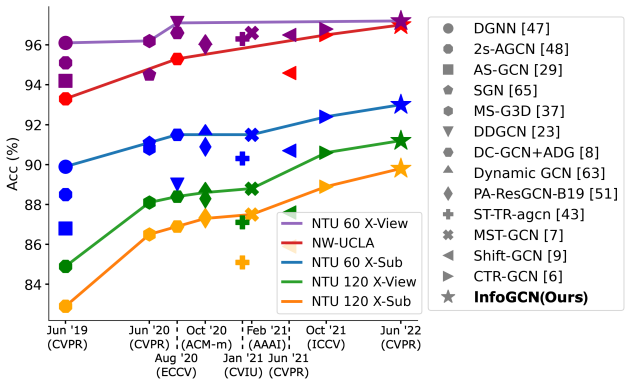

Figure 1. Leaderboard of skeleton based action recognition in recent two years. We connect the previous state-of-the-art records with lines, showing ours brings considerable gains over most datasets. An approach-by-approach comparison clearly shows our performance improvement.

D.2. Ablation Studies on NTU RGB+D 120

We provide additional ablation studies on NTU RGB+D 120 cross-subject split with joint information ($k = K$) as in the main paper unless otherwise stated. Bold figures indicate the best performance in all the tables in this section.

Effect of the Number of Heads In Table 4, we compare the performance of InfoGCN with the different numbers of heads (M) along with the number of model parameters. We observe that InfoGCN with three heads shows the best performance compared to others. Despite the increase in model capacity, models with more than three heads do not improve the performance. The number of parameters of 3-headed InfoGCN is 1.57M which is slightly larger (around 0.1M) than previous state-of-the-art methods [2, 9, 12].

Composition Operations for the Context-Dependent Topology. To explore the design of context-dependent topology in the SA-GC module, we compare three different composition operators ($+$, $\odot$, and $\otimes$) for combining shared-topology with self-attention map in Table 5. Here, $\otimes$ denotes broadcasted matrix multiplication, and $\odot$ indicates broadcasted element-wise multiplication. We see that element-wise multiplication achieves the highest performance in terms of classification accuracy, and results verify the effectiveness of the context-dependent topology design.

Coefficients for MMD Losses We run a grid search over the coefficients of loss in the range of $\lambda_1 \in \{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$ and $\lambda_2 \in \{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}\}$ as shown in Table 6. We see that coefficient $\lambda_1 = 1 \times 10^{-4}$ and $\lambda_2 = 1 \times 10^{-1}$ gives the best performance. We search the loss coefficients for the experiments on other datasets in the same way.

M	# Params	Acc (%)
1	1.03M	84.8
2	1.30M	84.8
3	1.57M	**85.1**
4	1.85M	84.7
5	2.12M	84.8
6	2.39M	84.5

Table 4. Accuracies of the different numbers of head of InfoGCN.

Methods	Acc (%)
$\tilde{\mathbf{A}} + \mathrm{SA}(\mathbf{H}_t)$	84.7
$\tilde{\mathbf{A}} \otimes \mathrm{SA}(\mathbf{H}_t)$	84.9
$\tilde{\mathbf{A}} \odot \mathrm{SA}(\mathbf{H}_t)$	**85.1**

Table 5. Accuracies of the different composition operations for context-dependent topology.

		λ_2		
		1×10^{-3}	1×10^{-2}	1×10^{-1}
λ_1	1×10^{-5}	84.9	84.8	84.9
	1×10^{-4}	84.5	84.7	**85.1**
	1×10^{-3}	84.9	84.6	84.8
	1×10^{-2}	85.0	84.8	84.7
	1×10^{-1}	84.7	84.6	84.6

Table 6. Accuracies of the different coefficients of MMD loss.

k	Acc (%)	
	Pos.	Mot.
1 (Bone)	**87.3**	**82.5**
2	86.5	82.2
3	85.3	82.2
4	84.4	81.8
5	84.7	82.2
6	84.7	81.9
7	84.6	82.0
8 (Joint)	85.1	82.1

Table 7. Accuracies of the different k-th mode representation of skeleton on NTU RGB+D 120 cross-subject split.

Multi-modal Ensemble To examine the effect of the multi-modal ensemble, we ensemble different sets of k-th mode representations of the skeleton. We present the action classification accuracy of each k-th mode representation in Table 7 and their ensemble in Table 8. The *k-set* column in Table 8 (and Tables 9 to 12) indicates a set of modalities used for the ensemble. The Pos. and Mot. in the tables indicates position and motion, respectively. The *Pos. & Mot.* column in the tables represents the ensemble of position and motion of skeleton joint features in the corresponding k-set. In Table 7, we see that the model trained with bone infor-

Modes	k-Set	Acc (%)		
		Pos.	Mot.	Pos.&Mot.
Bone	$\{1\}$	87.3	82.5	88.9
Joint	$\{K\}$	85.1	82.1	86.9
4 ensemble	$\{1,K\}$	88.5	84.2	89.4
6 ensemble	$\{1,2,K\}$	89.0	84.9	**89.8**
8 ensemble	$\{1,2,3,K\}$	89.0	84.9	89.6
10 ensemble	$\{1,2,3,4,K\}$	89.5	85.4	89.3
12 ensemble	$\{1,2,3,4,5,K\}$	89.4	85.5	89.3
14 ensemble	$\{1,2,3,4,5,6,K\}$	89.3	85.4	89.2
16 ensemble	$\{1,2,3,4,5,6,7,K\}$	89.2	85.5	89.2

Table 8. Multi-modal ensemble results on NTU RGB+D 120 cross-subject split. K is equal to 8.

Modes	k-Set	Acc (%)		
		Pos.	Mot.	Pos.&Mot.
Bone	$\{1\}$	90.6	88.6	92.2
Joint	$\{K\}$	89.8	88.9	91.2
4 ensemble	$\{1,K\}$	91.6	90.1	92.7
6 ensemble	$\{1,2,K\}$	92.2	90.6	**93.0**

Table 9. Multi-modal ensemble results on NTU RGB+D 60 cross-subject split. K is equal to 8.

Modes	k-Set	Acc (%)		
		Pos.	Mot.	Pos.&Mot.
Bone	$\{1\}$	95.5	93.6	96.2
Joint	$\{K\}$	95.2	94.2	96.4
4 ensemble	$\{1,K\}$	96.5	95.0	96.9
6 ensmeble	$\{1,2,K\}$	96.7	95.3	**97.1**

Table 10. Multi-modal ensemble results on NTU RGB+D 60 cross-view split. K is equal to 8.

mation achieves the best performance in both position and motion. We also note that the performance of the model trained with positions tends to decrease as the k value increases except when k is equal to 8 (joint). However, the performance of the model trained with the motion has similar values over k with the range of $[81.9, 82.2]$, while the model trained with bone information ($k = 1$) shows the best performance.

In Table 8, we further observe that the ensemble of models trained with different k-th mode representations improves the classification accuracy, but after ensembling 6 modes, the accuracy decreases as the number of modalities increases. We attribute this to the fact that the k-th mode closed to K such as $k = 7$ or 8 does not provide distinctive features for classifying action. Therefore, the more we ensemble models trained with the k-th mode close to K, the greater the influence of the model trained with the K-th mode. Therefore, classification accuracy decreases after ensembling 6 modes, converging to the performance of the model with K-th mode.

Modes	k-Set	Acc (%)		
		Pos.	Mot.	Pos.&Mot.
Bone	$\{1\}$	88.5	84.8	90.1
Joint	$\{K\}$	86.3	84.4	88.4
4 ensemble	$\{1,K\}$	89.7	86.5	90.7
6 ensemble	$\{1,2,K\}$	90.3	86.9	**91.2**

Table 11. Multi-modal ensemble results on NTU RGB+D 120 cross-setup split. K is equal to 8.

Modes	k-Set	Acc (%)		
		Pos.	Mot.	Pos.&Mot.
Bone	$\{1\}$	95.3	95.5	94.8
Joint	$\{K\}$	94.0	93.1	94.6
4 ensemble	$\{1,K\}$	96.3	95.7	96.6
6 ensemble	$\{1,2,K\}$	97.0	95.5	**97.0**

Table 12. Multi-modal ensemble results on NW-UCLA. K is equal to 6.

D.3. Multi-modal Ensemble on All Benchmarks

We provide extensive experimental results of the multi-modal ensemble of each dataset and split in Tables 8 to 12. The value K, which denotes joint information in k-th mode skeleton representation, is 8 for NTU RGB+D 60 & 120 and 6 for NW-UCLA. We note that the performance increases as the number of modalities for ensemble increases in all the datasets.

D.4. Further Analysis on Information Bottleneck Constraint

To illustrate the effectiveness of the information bottleneck objective and its corresponding loss, we provide additional cases of PCA analysis conducted on Sec. 5.1. We trained our model with or without MMD loss and compared latent representations of different classes using PCA as shown in Fig. 2. In detail, we randomly sampled five action classes for visual simplicity from NTU RGB+D 120 dataset and repeated this process for 12 times. We observe a similar tendency in Fig. 2 as in Fig. 4. The latent rep-

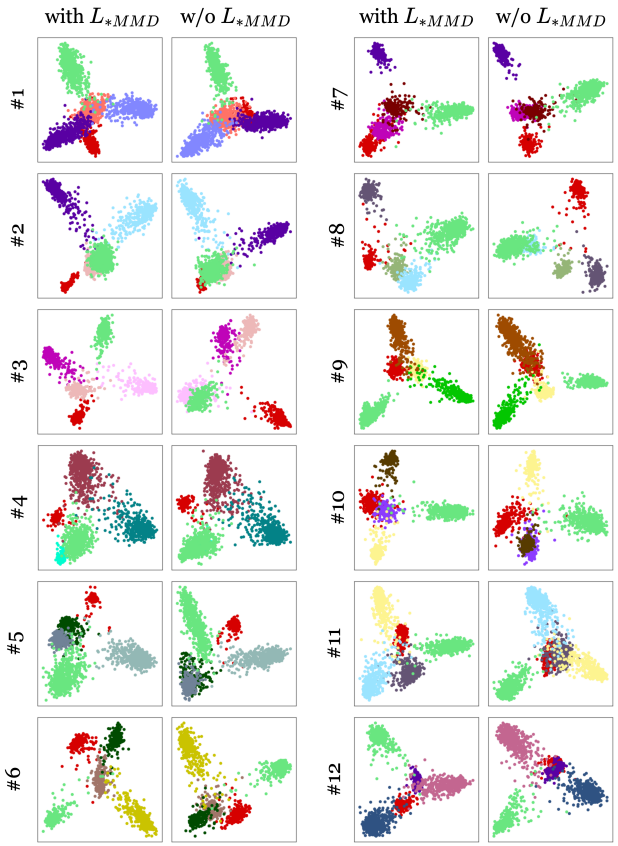

Figure 2. Various examples of PCA projection of latent representation to 2D when trained with or without MMD loss. We assign different colors for different action classes.

resentation learned with MMD loss shows less overlapping class conditional distributions that seem easier to discriminate than those without MMD loss.

D.5. Qualitative Results

To observe the patterns of how a joint attend others or to be attended, we illustrate examples of the skeleton and its corresponding topology inferred by InfoGCN in Figs. 3 to 7. Colored lines indicate inferred topology from a selected joint 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 as in the main paper. For convenience sake, we refer to the attention values of joint attending others as out-attention and those of other joints toward a joint as in-attention. The out-attention corresponds to rows of the self-attention map and the in-attention to columns.

In Fig. 3, we show minimal examples to compare four factors: in/out-attention, joint, time, and action class. We can see that the in-attention and out-attention differ for given time, action, and joint. As we observed from the self-attention maps in the main paper, the inferred topology has

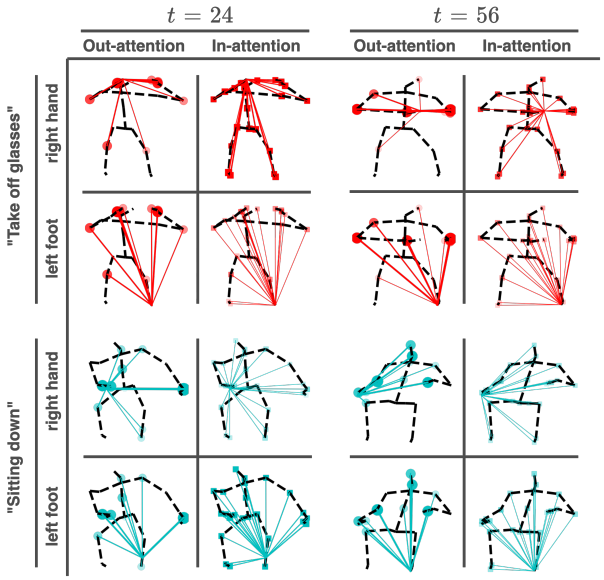

Figure 3. Examples of inferred context-dependent topology in different joints, times, and action classes. We visualize in-attention and out-attention separately.

asymmetric relation, unlike bone connectivity. The joint relations of inferred topology are varied depending on the direction. On the other hand, the attention patterns are similar when we compare the attention for different joints in the same in/out type, action, and time, but the average intensity differs.

To see the patterns adhere across joints, actions, and times, we present skeleton and inferred topology for different types of attention, joints, time frames, and action classes in Figs. 4 to 7. We choose representative joints (limps, head, the center of mass) and randomly select 24 different action classes from the test set on NTU RGB+D 120 dataset cross-subject split. We uniformly sample frames of each sequence of skeletons every 16 timesteps. The following is a list of the characteristics of the graph inferred by InfoGCN that we observe.

- The graph has asymmetric relations so that two different joints can have various relations depending on the direction.
- The in-attentions or out-attentions of the graph in specific time and action tend to be similar against entire joints.
- Graphs are dynamically inferred over time while dependent on the action, which helps recognize human behavior using the discriminative information provided by SA-GC.

Figure 4. Illustrations of context-dependent intrinsic topology (1/4)

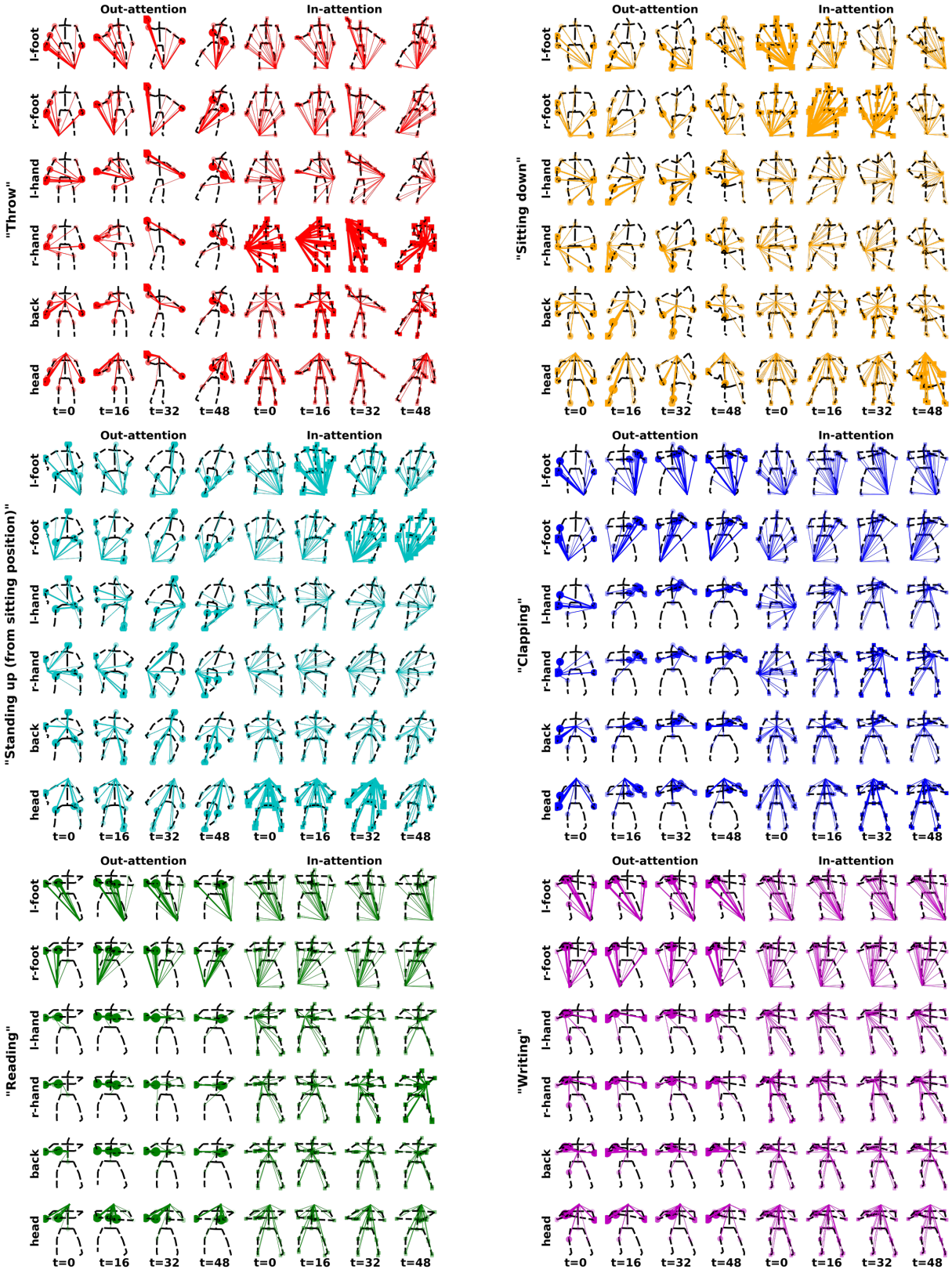

Figure 5. Illustrations of context-dependent intrinsic topology (2/4)

Figure 6. Illustrations of context-dependent intrinsic topology (3/4)

Figure 7. Illustrations of context-dependent intrinsic topology (4/4)

References

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

[2] 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. 4, 5

[3] 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. 4

[4] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015. 2

[5] Ian Fischer. The conditional entropy bottleneck. *Entropy*, 22(9):999, 2020. 1, 2

[6] 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. 2

[7] 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. 4

[8] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 2

[9] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8561–8568, 2019. 5

[10] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015. 2

[11] 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. 3, 4

[12] 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. 5

[13] Alireza Makhzani and Brendan Frey. Pixelgan autoencoders. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1972–1982, 2017. 2

[14] 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. 4

[15] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 4

[16] 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. 3, 4

[17] 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. 4