How helpful are current graph models for skeleton-based action recognition? A topology agnostic approach

Anonymous authors
Paper under double-blind review

Abstract

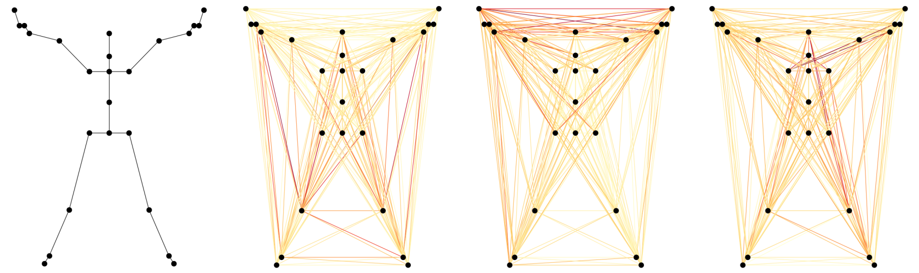
Graph Convolutional Networks (GCNs) have been dominating skeleton-based action recognition in recent years. While GCN-based approaches keep establishing new state-of-the-art results, the proposed architectures are getting increasingly sophisticated with a variety of add-ons. Many recent works attempt to relax the topology restriction imposed by the GCN framework, such as local/sparse connections and permutation invariance. However, the room for further innovation is extremely limited under such a framework. In this work, we present Topology-Agnostic Network (ToANet), a simple architecture based merely on Fully-Connected (FC) layers, as opposed to GCNs for skeleton-based action recognition. It is constructed by chaining FC layers applied across joints (aggregate joint information) and within each joint (transform joint features) in an alternate manner. Moreover, it contains a novel design of parallel paths for multi-relational modeling. ToANet proves to be a powerful architecture for learning the joint co-occurrence of human skeleton data. ToANet achieves better or comparable results to state-of-the-art GCNs on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA datasets. These results challenge the convention of choosing GCNs as the de-facto option for skeleton-based action recognition. We hope that our work stimulates further research on non-GCN based methods, eliminating the restriction of topology.

1 Introduction

Human action recognition has a broad range of real-world applications, such as video surveillance and human-machine interaction. In recent years, skeleton-based human action recognition methods have gained increased popularity due to their robustness to variations in lighting conditions, camera viewpoints, etc. The skeleton data can also be easily acquired by depth sensors or pose estimation algorithms Cao et al. (2017).

Earlier approaches construct a sequence of features or a pseudo-image from human joints as input for Recurrent Neural Networks (RNNs) Liu et al. (2016); Li et al. (2018) or Convolutional Neural Networks (CNNs) Ke et al. (2017); Liu et al. (2017) to generate the prediction. Nevertheless, these methods hardly capture the inherent correlations between human joints, which are intuitively crucial for human action recognition. For example, the action 'drinking' needs to be accomplished by arm, hand and head together. Yan et al. (2018) first propose to treat joints and their physical connections as nodes and edges of a graph, then a GCN is employed on such a predefined graph to learn the joint interactions.

In spite of its significant improvement on previous RNN- or CNN-based approaches, the manually defined topology ignores the relationship between physically unconnected joints, thus limiting the representation power of GCNs. Moreover, the hierarchical structure of GCNs is supposed to capture multilevel semantic information and predefined connections may not be suitable for higher layers. To address this issue, most recent approaches Gao et al. (2022); Cheng et al. (2020); Shi et al. (2019); Ye et al. (2020); Chen et al. (2021); Song et al. (2021); Xia & Gao (2021) employ a learnable topology and merely initialize it as the natural one imposed by the skeleton. After training, the learned topology is indeed fully connected. Many among them also add an adaptive component via



(a) Joints and their physical connections of skeleton data.

(b) Relation 1: focus on co-occurrence related to legs.

(c) Relation 2: focus on co-occurrence between both hands.

(d) Relation 3: focus on co-occurrence related to head and shoulder.

Figure 1: Visualization of the learned inter-joint FC layer's weights of our ToANet trained on the NTU RGB+D 120 dataset. Darker lines (the color ranges from shallow yellow to dark red with increased weights) indicates larger weights. It can be seen that ToANet automatically learns to focus on relations with a specific group of joints at each path. Many of the joint pairs with large weights are indeed those that interact intensively with each other.

attention or similar mechanisms, which are also fully connected but with dynamic weights instead. Such fully-connected topologies, including learnable and dynamic connections, has almost become the default choice of recent works and it contributes significantly to the improved performance over previous GCN-based models, as experimentally shown, e.g. in Shi et al. (2019).

These results conform to the human intuition that considering the relation between all joint pairs, regardless of whether they are physically connected or not, is necessary for better performance in skeleton-based human action recognition. Consequently, the concept of topology is of little significance in such fully-connected models Chen et al. (2021), except for initialization. Yet, in our analysis (see Sec. 4.2), we also see that the initialization is redundant. Therefore, we deduce that an elementary fully-connected layer applied on the spatial dimension is theoretically adequate for modelling the relation between all joint pairs, by simply stacking them along the spatial dimension.

Motivated by these observations, we propose in this paper a purely FC-layer-based model for the spatial modeling of skeleton-based human action recognition, called ToANet. It chains a range of FC layers applied on joint and feature dimension alternately. The former is intended to aggregate global information from other joints, whereas the latter plays the role of feature projection. Based on this architecture, we further propose a novel design for learning multi-relational interactions between joints. Semantically, there may exist different types of relationships between joints. For example, wrists and elbows both belong to the upper limbs, but they also cooperate intensively with far-away joints, e.g., ankles for actions such as ”jumping”, etc. Therefore, we assume that the consideration of multiple relations contributes to a more realistic modeling of the skeleton data.

There has been an on-going trend of exploring architectures with less hard-coded priors so that higher capacity and flexibility can be provided, e.g., CNNs and GCNs had saved researchers from the heavy routine of handcrafting features and had created more room for architecture designs. Recently, vision transformers Dosovitskiy et al. (2021); Liu et al. (2021) and MLP-based models Tolstikhin et al. (2021); Touvron et al. (2021) further get rid of the convolutional inductive bias for image data, namely local/sparse connectivity and translation invariance. Analogously, our proposed topology-agnostic network provides a more flexible model for skeleton-based human action recognition, which we hope may stimulate follow-up studies without being restricted to topology modelling in the current research landscape of skeleton-based human action recognition.

In summary, we contribute, first, a model purely based on fully connected layers for the spatial modeling of skeleton-based human action recognition, called ToANet. Second, we propose a novel design for learning multi-relational interactions between human body joints. We further show that the topology-agnostic modeling with fully connected layers we propose can reach or exceed results on par with recent GCN based models on the commonly used datasets NTU RGB-D Shahroudy et al. (2016), NTU RGB-D 120 Liu et al. (2019) and Northwestern-UCLA Wang et al. (2014).

2 RELATED WORK

2.1 SKELETON-BASED HUMAN ACTION RECOGNITION

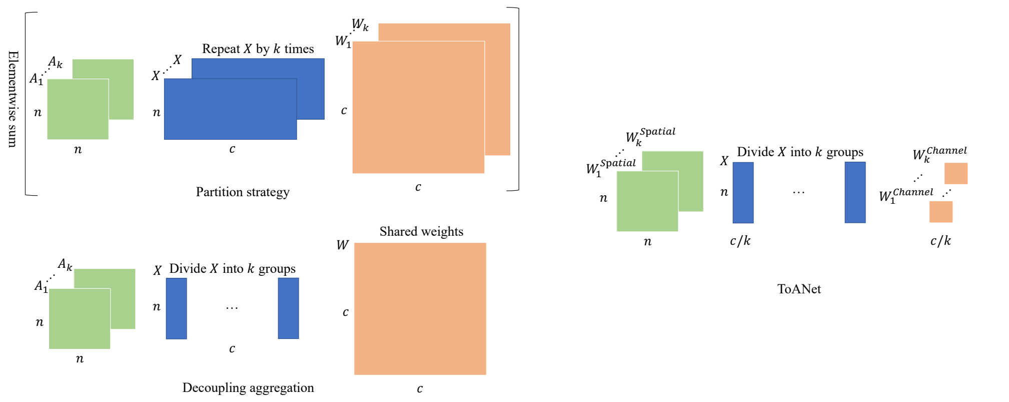
Considering deep neural network approaches, RNNs Du et al. (2015); Song et al. (2017); Zhang et al. (2017) have been a first popular choice to tackle skeleton-based human action recognition. The application of CNNs for this task is also well studied Ke et al. (2017); Liu et al. (2017). Yet, the spatial interactions of body joints are not explicitly given in the such methods. In contrast, GCNs can model the spatial configurations of joints as a graph. We will focus our following review on such graph-based models since they have become the de-facto choice for this task.

2.2 GCN-BASED SKELETON ACTION RECOGNITION

Graph convolution is a generalization of the convolution operation in the image space to the non-Euclidean space. To apply convolution on graphs, there is an additional challenge of how to handle different numbers of neighboring nodes. Comparing to many advanced GCN models such as GIN Xu et al. (2018) and MPNN Gilmer et al. (2017), the GCN proposed by Kipf & Welling (2016) is widely adopted for action recognition due to its simplicity and thus higher resistance to overfitting. It first applies a spatially-shared feature transformation on them and then aggregates information of neighboring nodes by a weighted sum of their transformed features (the order is invertible according to associative law of matrix multiplication). Finally, a scalar nonlinearity is added to the end.

Yan et al. (2018) first introduce a GCN to model the joint correlations and demonstrate its effectiveness for action recognition. For GCNs, the topology defines the vertex connectivity and thus plays a crucial role. Yan et al. (2018) simply assume a fixed topology according to the natural connections of joints. However, its limitation is identified later, and many follow-up works focus on the **topology** of GCNs for action recognition. The topology of recent proposed GCNs falls into one or more of the following four categories:

- **Learnable Topology** Most recent approaches Gao et al. (2022); Cheng et al. (2020); Shi et al. (2019); Ye et al. (2020); Chen et al. (2021); Song et al. (2021); Xia & Gao (2021); Liu et al. (2020); Chi et al. (2022) are based on a learnable topology which considers the relationship between both physically connected and unconnected joints.
- **Dynamic Topology** Many among them Gao et al. (2022); Cheng et al. (2020); Shi et al. (2019); Chen et al. (2021); Ye et al. (2020); Chi et al. (2022) also adopt attention or similar mechanisms to produce a data-dependent component of the topology (analogous to Graph Attention Networks Veličković et al. (2017)), boosting GCN's performance further. Chi et al. (2022) propose an novel loss to improve over Chen et al. (2021), which they build upon.
- **Channel-Specific Topology** Both CTR-GCN Chen et al. (2021) and Decoupling GCN Cheng et al. (2020) propose to consider the variation of topology along the channel dimension in a similar fashion, i.e. dividing the channels into multiple groups (up to number of channels) with independent or refined topologies. This technique can increase the expressiveness of spatial aggregation to some extent without adding Floating Point Operations (FLOPs).
- **Spatially Partitioned Topology** For images, the weight function varies at different spatial positions of the convolution kernel. For instance, the weight $W \in \mathbb{R}^{C,K,K}$ of a $K \times K$ 2D convolution kernel contains $K \times K$ slices of unique weight vectors for each spatial position. To adapt convolution for graph data, the adjacency matrix is often partitioned into multiple subsets Niepert et al. (2016); Yan et al. (2018). Each subset is then assigned a corresponding convolution weight vector. Manually designing and validating a complex partition strategy is laborious, and the computation of each subset is not parallelizable. So the number of partitions is typically less than three. For skeleton-based human action recognition, a practically efficient and thus widely adopted strategy Yan et al. (2018); Chi et al. (2022); Chen et al. (2021); Shi et al. (2019); Cheng et al. (2020); Zhang et al. (2020); Liu et al. (2020) is the so-called spatial configuration partitioning Yan et al. (2018), which divides the adjacency matrix into three matrices according to the distance from the referenced node.



(a) Multi-relational modeling methods for GCNs.

(b) Our ToANet for multi-relational modeling.

Figure 2: Compare our method to other multi-relational modelling approaches including decoupling aggregation Cheng et al. (2020) and the default partition strategy Yan et al. (2018) in recent GCN-based approaches.

3 Method

3.1 Revisiting GCNs for Skeleton-based Human Action Recognition

The original GCN layer has the formulation Kipf & Welling (2016)

$$H^{(l+1)} = \sigma(\tilde{A}H^{(l)}W^{(l)}), \tag{1}$$

where $\tilde{A} \in \mathbb{R}^{V \times V}$ denotes the normalized adjacency matrix, $H \in \mathbb{R}^{V \times T \times C}$ and $W \in \mathbb{R}^{C \times C}$ are hidden representation and weight matrix applied on the channel dimension, respectively, where V denotes for the number of joints, T the number of frames and Cdenotes the number of channels. σ is the nonlinear ReLU activation and the superscript l indicates layer number. State-of-the-art GCN-based approaches make adaptations to the above formulation to arrive at high accuracy predictions. We summarize the characteristics of the topologies defined by recently proposed GCN-based methods as follows:

- **Fully-Connected Modeling** Notably, both the **learnable** and **dynamic** topologies in Cheng et al. (2020); Shi et al. (2019); Ye et al. (2020); Chen et al. (2021); Song et al. (2021); Xia & Gao (2021); Liu et al. (2020) are assuming a fully-connected topology. It is also intuitively better to consider the semantic connections between all joint pairs instead of only the physically linked joints, challenging the need for GCNs in this context.
- **Multi-Relational Modeling** We find that both the **channel-specific** Cheng et al. (2020); Chen et al. (2021) and **spatially partitioned** topologies are essentially different ways of modelling a multi-relational graph. For learnable topologies with/without a dynamic component, those handcrafted partition rules actually do not hold, because each subset becomes fully-connected after learning. In such a case, employing multiple partitions is equivalent to ensembling multiple paths at each layer. Channel-specific topologies can be regarded as another way of modelling such a multi-relational graph by dividing the channel dimension into multiple groups, instead of replicating the channel dimension multiple times. However, there is a gap between this technique and partitioning: it assigns the same convolution weight to different groups of topologies, whereas each topology is assigned a different convolution weight for the partitioning approach, see Fig. 2.

3.2 Our Proposed ToANet

Assuming all the joints are (semantically) connected to one another, the topology is no longer important and we can directly stack joints together to form the spatial dimension of the input. Equivalently, this can be achieved by simply replacing the adjacency matrix A with a weight matrix

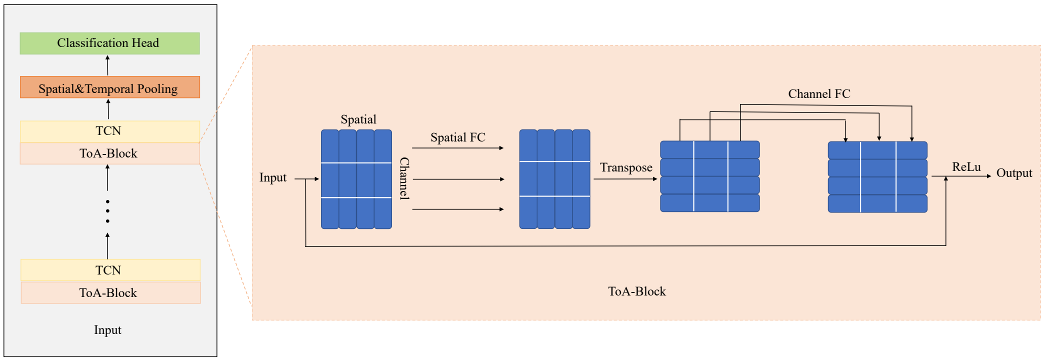


Figure 3: An overview the proposed ToANet with temporal convolution modules. The basic block of ToANet consists of two fully-connected layers applied alternately on the spatial and the channel dimension. In addition, the fully-connected layers are divided into multiple parallel paths, in order to model the multi-relational joint co-occurrence and interdependence.

$W_{spatial} \in \mathbb{R}^{V \times V}$ applied on the spatial dimension in Eq. (1):

$$H^{(l+1)} = \sigma(W_{spatial}^{(l)} H^{(l)} W^{(l)}). \tag{2}$$

Surprisingly, this results in an architecture which resembles ResMLP Touvron et al. (2021) and MLP-Mixer Tolstikhin et al. (2021) for image data, except that it is based on single fc layers instead of MLPs.

3.2.1 Multi-Relational Modeling

Inspired by multi-head self-attention Vaswani et al. (2017), we propose a novel approach to model the multi-relational spatial configuration of skeleton data. As shown in Fig. 3, the feature dimension is divided into K paths ($K = 3$ for illustration), then both spatial aggregation and feature projection are applied on each k^{th} path in parallel:

$$H_k^{(l+1)} = \sigma(W_{k,spatial}^{(l)} H_k^{(l)} W_k^{(l)}) \tag{3}$$

where $H_k \in \mathbb{R}^{V,T,C/K}$ and $W_k \in \mathbb{R}^{C/K \times C/K}$. Our approach differs from the decoupling aggregation Cheng et al. (2020) in two aspects:

- At an operation level, our method assigns a unique weight vector to each group, whereas the decoupling method assumes a shared weight vector.
- The decoupling method is intended to bridge the gap between convolution on graphs and images, whereas our method aims at multi-relational modeling of skeleton data.

3.2.2 Architecture

As shown in Fig. 3, the whole network consists of a number of elementary blocks, followed by a spatial-temporal global average pooling layer and a linear layer to predict action classes. There are a total of three stages and the number of channels are doubled at the beginning of each stage. Following Chen et al. (2021); Liu et al. (2020); Chi et al. (2022), we adopt a multi-scale temporal convolution module with 4 branches, including two temporal convolutions with different dilations of 1 and 2, Max Pooling as well as residual connection respectively. A 1x1 convolution is also applied at the beginning of each branch for reducing channel dimension. The stride is 1 by default, except that they are set to 2 at the beginning of each stage. The basic block of our ToANet is illustrated in Fig. 3. Each block is composed of a spatial module, i.e. our ToANet, and a temporal convolution module.

3.2.3 DIFFERENCES TO GCN-BASED APPROACHES

Our architecture differs from recent proposed GCNs with the following simplifications:

- **Topology-Agnostic** As discussed in Sec. 2, recent GCNs are already utilizing a fully-connected topology. However, they stick to the physical connections for initialization. More importantly, many architecture designs are restricted by the concept of topology. For example, the study of partition strategies in Yan et al. (2018), channel-wise topology refinement in Chen et al. (2021), etc.. Our simpler model offers a freedom for variations.
- **Attention-Free** Most modern GCNs are entangled with dynamic connections between joints via attention or similar mechanisms, whereas our pure fully-connected model achieves state-of-the-art results. Thus our ToANet serves as a clean baseline for further development. It is likely that our model can be the basis of even better performing approaches, when equipped with attention or similar mechanisms.

4 EXPERIMENTS

In this section, we first analyze the role of topology information in GCN-based models for action recognition, following our intuition that topology actually plays a minor role in these models. Then, we conduct an ablation study for a deeper understanding of the proposed ToANet. Finally, we compare ToANet to other state-of-the-art approaches on skeleton-based human action recognition benchmarks and show results on par with the state of the art Cheng et al. (2020); Chi et al. (2022). While Chi et al. (2022) show that further improvements can be achieved by adopting an improved loss, we base our experiments on the standard cross entropy loss. This allows for an evaluation of our architecture as such and for direct comparability to most previous work.

4.1 SETTINGS

We evaluate ToANet on three commonly used benchmark datasets NTU-RGB+D Shahroudy et al. (2016), NTU-RGB+D120 Liu et al. (2019) and Northwestern-UCLA Wang et al. (2014).

4.1.1 DATASETS

NTU RGB+D NTU-RGB+D Shahroudy et al. (2016) contains 56880 samples which are conducted by 40 volunteers and categorized into 60 classes. The skeleton-data is captured by three Microsoft Kinect v2 depth sensors from different horizontal views. The authors suggest two benchmarks:

- Cross-subject (X-sub): the subjects in the training and test subsets are different.
- Cross-view (X-view): the training set contains 37920 samples captured by the sensors at $0°$ and $45°$, and the testing set includes 18960 sequences captured by the sensor at $-45°$.

NTU RGB+D 120 NTU-RGB+D120 Liu et al. (2019) is extended from NTURGB+D with extra 57367 skeleton sequences over 60 additional action classes. It contains 32 setups, each of which denotes a specific location and background. The authors suggest two evaluation protocols:

- Cross-subject (X-sub): the subjects in the training and test subsets are different.
- Cross-setup (X-setup): the samples in the training and test subsets have different setup IDs.

UCLA Northwestern-UCLA Wang et al. (2014) dataset is captured by three Kinect cameras from different viewpoints. Following Wang et al. (2014): we construct the training data using samples from the first two cameras, and the testing data using samples from the other camera.

4.1.2 IMPLEMENTATION

All experiments are conducted on a single Tesla V100 GPU with the PyTorch Paszke et al. (2019) deep learning framework. A total number of 110 epochs is chosen for all the experiments and the warmup method in He et al. (2016) is adopted in the first 5 epochs for a more stable training process. We train the model using Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9) and

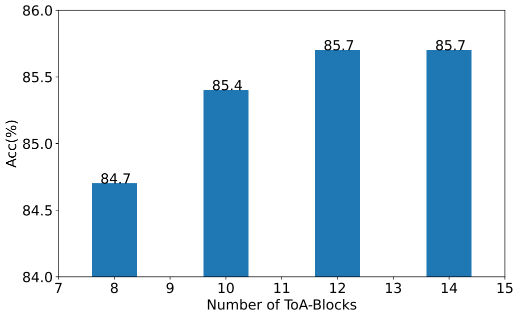


Figure 4: Effect of the number of layers.

weight decay (0.0004 for NTU RGB+D and NTU RGB+D 120, 0.0001 for Northwestern-UCLA) for optimization. We apply cross-entropy loss as the objective function. The learning rate is initialized to 0.1 for NTU RGB+D and NTU RGB+D 120 and is reduced by a factor of 10 at epoch 90 and 100, following Chi et al. (2022). For Northwestern-UCLA, we adopt a smaller learning rate of 0.05 and the same decay schedule. For NTU RGB+D and NTU RGB+D 120, the batch size is set to 64, each sample is resized to 64 frames, and we follow the data pre-processing in Zhang et al. (2020). For Northwestern-UCLA, we use a batch size of 16, and adopt the data pre-processing as in Cheng et al. (2020); Chen et al. (2021). Our code is based on the official implementation of Chen et al. (2021) and Zhang et al. (2020).

4.2 Preliminaries

The inspiration of ToANet is drawn from the following experiments, analysing the role of topology.

Initialization of Adjacency Matrix	Acc(%)
Physical Connections Chen et al. (2021)	83.9
Identity Matrix	**84.0**
Ones	83.8
Normal Distribution	83.6

Table 1: The effect of the initialization of adjacency matrix on the X-sub benchmark of NTU RGB+D 120. The GCN baseline proposed by Chen et al. (2021) is used for comparison. Note that our reproduced result of 83.9 is indeed higher than their reported 83.7.

To validate our analysis that the concept of topology is of little significance for GCN-based methods, we compare different ways of initializing the adjacency matrix, from the special initialization using physical connections as in Chen et al. (2021) to topology agnostic ones. For this experiment, we take a strong baseline model proposed in Chen et al. (2021) which established top performance on the X-sub benchmark of NTU RGB+D 120, using basic GCN layers with a learnable topology. The experimental setup except for the initialization is kept exactly the same as in Chen et al. (2021). Our results indicate that the special initialization of the adjacency matrix according to physical connections is not needed for recent GCN models with learnable topologies.

4.3 Ablation Analysis

In this section, we compare our approach with other methods designed for multi-relation modeling. For ablation study, all the experiments are conducted on the X-sub setup of NTU RGB+D 120 using a single modality of joint coordinates as input.

Method	Groups	Channels	layers	Parameters	Acc(%)
Fully-Connected Baseline (FCB)	1	128	12	3.7M	84.9
FCB (Ours)	3	128	12	6.0 M	85.5
+ Partition Yan et al. (2018)	6	128	12	9.4M	85.3
FCB (Ours) + Decoupling Cheng et al. (2020)	4	128	12	3.7M	84.8
	8	128	12	3.7M	85.2
	16	128	12	3.8M	85.3
ToANet (ours)	4	128	12	2.1M	85.4
	8	128	12	2.0M	**85.7** (+0.8)
	16	128	12	2.0M	85.3

Table 2: Ablation of multi-relational approaches on the X-sub benchmark of NTU RGB+D 120. For a fair comparison, we adapt the GCN-based methods including Spatial Configuration Partitioning Yan et al. (2018) and Decoupling Aggregation Cheng et al. (2020) for our Fully-Connected Baseline.

4.3.1 Comparing to Other Multi-relation Modeling Methods

As the core feature of our proposed ToANet, its multi-relation modeling design raises the classification accuracy by an absolute percentage of 0.8% over the Fully-Connected Baseline, with significantly reduced parameters. To show the effectiveness of our ToANet, we compare to other approaches including Spatial Configuration Partitioning Yan et al. (2018) and Decoupling Aggregation Cheng et al. (2020) for modeling multi-relational data (see Sec. 2 and Sec. 3.1). Moreover, we apply these two techniques to our Fully-Connected Baseline for a fair comparison. From Tab. 2, we see that ToANet achieves the best performance, outperforming Decoupling Aggregation by 0.4% and Spatial Configuration Partitioning by 0.2%. In addition, the partition strategy brings extra parameters and memory usage, as opposed to the Decoupling Aggregation and our method ToANet.

4.4 Visualization of Learned Weights

We have visualized the learned weights of our ToANet at the 12^th spatial FC layer in Fig. 1. It can be seen that ToANet automatically learns to focus on relations with a specific group of joints at each path. Many of the joint pairs with large weights are indeed those which interact intensively with one another. This validates our intuition that joint interactions can be learned by topology-agnostic architectures and explains the excellent performance of our ToANet.

4.5 Comparison to the State-of-The-Art

To utilize the complementary information between different modalities, many state-of-the-art approaches employ a multi-stream fusion strategy. For a fair comparison, we follow the same multi-stream fusion strategy as Cheng et al. (2020); Ye et al. (2020); Chen et al. (2021); Chen et al. (2021), i.e., we utilize 4 streams taking joint, bone, joint motion and bone motion as input respectively. Joint refers to the original skeleton coordinates, and bone is represented by the relative coordinates between all naturally connected joint pairs. The joint motion and bone motion takes the difference between two temporally adjacent frames of corresponding data as input. The fusion is simply applied on the final scores of those four streams.

The comparison is conducted on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA in Tab. 3 and Tab. 4 respectively. Note that the recently published Chi et al. (2022) and Duan et al. (2022) are not directly comparable to our method. Duan et al. (2022) improve results using additional RGB input which requires heavy computation. InfoGCN Chi et al. (2022) proposed an extra loss which is orthogonal to architecture design. We can however compare to their model without this loss referring to the results they report in the main text for NTU-RGB+D 120. Yet, even this number refers to the ensemble of 6 modalities, which provides an additional advantage over the 4 modalities used in other models including ours.

Our ToANet exceeds 89% accuracy on the most challenging NTU-RGB+D 120 Cross-Subject benchmark in Tab. 3, improving over Cheng et al. (2020) and Chi et al. (2022) with the novel loss. It performs slightly worse than the state-of-the-art GL-CVFD Gao et al. (2022) on the smaller dataset

Method	NTU-RGB+D		NTU-RGB+D 120	
	X-Sub(%)	X-View(%)	X-Sub(%)	X-Set(%)
VA-LSTM Zhang et al. (2017)	87.7	79.2	-	-
2s-AGCN Shi et al. (2019)	88.5	95.1	82.9	84.9
4s-shift-GCN Cheng et al. (2020)	90.7	96.5	85.9	87.6
DC-GCN+ADG Cheng et al. (2020)	90.8	96.6	86.5	88.1
MS-G3D Liu et al. (2020)	91.5	96.2	86.9	88.4
PA-ResGCN-B19 Song et al. (2020)	90.9	96.0	87.3	88.3
Dynamic GCN Ye et al. (2020)	91.5	96.0	87.3	88.6
MST-GCN Chen et al. (2021)	91.5	96.6	87.5	88.8
EfficientGCN-B4 Song et al. (2021)	91.7	95.7	88.3	89.1
CTR-GCN Chen et al. (2021)	92.4	96.8	88.9	90.6
Ta-CNN+ Xu et al. (2022)	90.7	95.1	85.7	87.3
InfoGCN Chi et al. (2022) w/o (w/) MMD losses	- (92.7)	- (96.9)	89.1 (89.4)	- (90.7)
GL-CVFD Gao et al. (2022)	**92.4**	**97.1**	88.9	**90.8**
ToANet	92.3	96.5	**89.2**	90.6

Table 3: Classification results of our ToANet and state-of-the-art methods on NTU RGB+D and NTU RGB+D 120.

NTU-RGB+D 60. However, note that GL-CVFD Gao et al. (2022) is three times larger than our model (6.5M vs. 2.0M) and they rely on a two-stage training strategy which requires much heavier computation. Keep in mind that many GCN-based approaches including CTR-GCN and InfoGCN also rely on mechanisms generating dynamic weights (attention or similar), whereas our ToANet is purely based on FC layers.

The Northwestern-UCLA dataset is particularly challenging since it only contains few training sequences. Thus, prior knowledge on the topology might have a bigger impact here. In comparison to Ta-CNN+ Xu et al. (2022), our model performs slightly worse on Northwestern-UCLA but significantly better on NTURGB+D and NTURGB+D 120. In comparison, our model is much smaller which fits the small datasets better, while Xu et al. (2022) can leverage the additional capacity on larger datasets. Moreover, Ta-CNN+ has used their proposed data augmentation technique called SkeletonMix..

Method	Northwestern-UCLA
	Top-1 (%)
Ensemble TS-LSTM Lee et al. (2017)	89.2
2s-AGC-LSTM Si et al. (2019)	93.3
4s-shift-GCN Cheng et al. (2020)	94.6
DC-GCN+ADG Cheng et al. (2020)	95.3
CTR-GCN Chen et al. (2021)	96.5
Ta-CNN+ Xu et al. (2022)	**97.2**
InfoGCN w/o (w/) MMD losses	- (96.6)
ToANet	96.6

Table 4: Classification results of our ToANet and state-of-the-art methods on Northwestern-UCLA.

5 Conclusion

In this paper, we revisit the concept of topology in skeleton-based human action recognition. With ToANet, we propose the first topology-agnostic model purely based on fully-connected layers that is able to compete with state-of-the-art results by GCNs on skeleton-based human action recognition benchmarks. Theoretically, this gives us a better understanding of the potential of models with less inductive biases. More importantly, we believe these results open new possibilities in this field, beyond the limitations of established GCN-based methods.

Reproducibility Statement To ensure reproducibility, we provide all training details and hyperparameters in Appendix B as well as the code for our model in the supplementary material.

References

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1302–1310, 2017.

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*, pp. 13359–13368, 2021.

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, pp. 1113–1122, 2021.

Ke Cheng, Yifan Zhang, Congqi Cao, Lei Shi, Jian Cheng, and Hanqing Lu. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *European Conference on Computer Vision*, pp. 536–553, 2020.

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*, pp. 183–192, 2020.

Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20186–20196, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021.

Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, 2015.

Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2969–2978, 2022.

Lingling Gao, Yanli Ji, Yang Yang, and HengTao Shen. Global-local cross-view fisher discrimination for view-invariant action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5255–5264, 2022.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pp. 1263–1272, 2017.

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*, pp. 770–778, 2016.

Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4570–4579, 2017.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR (Poster)*, 2016.

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*, pp. 1012–1020, 2017.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5457–5466, 2018.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pp. 816–833. Springer, 2016.

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.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

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*, pp. 143–152, 2020.

Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, pp. 2014–2023. PMLR, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8026–8037, 2019.

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*, pp. 1010–1019, 2016.

Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12026–12035, 2019.

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*, pp. 1227–1236, 2019.

Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

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*, pp. 1625–1633, 2020.

Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *arXiv preprint arXiv:2106.15125*, 2021.

Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021.

Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021.

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*, pp. 5998–6008, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

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*, pp. 2649–2656, 2014.

Hailun Xia and Xinkai Gao. Multi-scale mixed dense graph convolution network for skeleton-based action recognition. *IEEE Access*, 9:36475–36484, 2021.

Kailin Xu, Fanfan Ye, Qiaoyong Zhong, and Di Xie. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2866–2874, 2022.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks. In *International Conference on Learning Representations*, 2018.

Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, pp. 7444–7452, 2018.

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*, pp. 55–63, 2020.

Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2117–2126, 2017.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1112–1121, 2020.

A Broader Impact

The study of skeleton-based human action recognition is of great practical significance. It is not only computationally more efficient to use skeletons instead of raw videos, but it also resolves the special concern for privacy in the applications of human action recognition. For example, our model can be deployed for violence detection, at the same time keeping the crowds' identities anonymous.

B Reproducibility - Experiment Details

In order to ensure reproducibility, we provide all training hyperparameters for our method for all datasets in the following.

All experiments are conducted on a single Tesla V100 GPU with the PyTorch deep learning framework. A total number of 110 epochs is chosen for all the experiments and the warmup method in He et al. (2016) is adopted in the first 5 epochs for a more stable training process. We train the model using Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9) and weight decay (0.0004 for NTU RGB+D and NTU RGB+D 120, 0.0001 for Northwestern-UCLA) for optimization. We apply cross-entropy loss as the objective function. The learning rate is initialized to 0.1 for NTU RGB+D and NTU RGB+D 120 and is reduced by a factor of 10 at epoch 90 and 100, following InfoGCN Chi et al. (2022). For Northwestern-UCLA, we adopt a smaller learning rate of 0.05 and the same decay schedule. For NTU RGB+D and NTU RGB+D 120, the batch size is set to 64, each sample is resized to 64 frames, and we follow the data pre-processing in Zhang et al. (2020). For Northwestern-UCLA, we use a batch size of 16, and adopt the data pre-processing as in Cheng et al. (2020); Chen et al. (2021). Our code is based on the official implementation of Chen et al. (2021) and Zhang et al. (2020) and will be fully released upon acceptance.

We show in Tab. 5 the default hyperparameters for training our ToANet on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA.

Config.	NTU RGB+D and NTU RGB+D 120	Northwestern-UCLA
random choose	False	True
random rotation	True	False
window size	64	52
weight decay	4e-4	1e-4
base lr	0.1	0.05
lr decay rate	0.1	0.1
lr decay epoch	90, 100	90, 100
warm up epoch	5	5
batch size	64	16
num. epochs	110	110
optimizer	Nesterov Accelerated Gradient	Nesterov Accelerated Gradient

Table 5: Default hyperparameters for our ToANet on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA.

C More Experiment Results

We also provide the experiment results for each modality on different benchmarks in detail, see Tab. 6 and Tab. 7.

Modality	NTU-RGB+D 120		NTU-RGB+D	
	X-Sub(%)	X-Set(%)	X-Sub(%)	X-View(%)
Joint	85.7	87.3	90.2	94.6
Bone	86.6	88.6	90.4	95.6
Motion	82.7	84.2	88.2	92.9
Bone Motion	82.6	84.1	88.1	92.5

Table 6: Classification accuracy of our ToANet using different modalities on the NTU RGB+D and NTU RGB+D 120 dataset.

Modality	Northwestern-UCLA (%)
Joint	93.3
Bone	92.0
Motion	92.0
Bone Motion	90.3

Table 7: Classification accuracy of our ToANet using different modalities on the Northwestern-UCLA dataset.