Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training

Zhenjie Chen*
School of Cyber Science and Engineering, Southeast University
Engineering Research Center of Blockchain Application, Supervision And Management
Nanjing, China
chenzhenjie@seu.edu.cn

Hongsong Wang* †
Department of Computer Science and Engineering, Southeast University
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications
Nanjing, China
hongsongwang@seu.edu.cn

Jie Gui†
School of Cyber Science and Engineering, Southeast University
Purple Mountain Laboratories
Engineering Research Center of Blockchain Application, Supervision And Management
Nanjing, China
guijie@seu.edu.cn

ABSTRACT

Recently, skeleton-based human action recognition has received widespread attention in computer vision community. However, most existing research focuses on improving the recognition accuracy on complete skeleton data, while ignoring the performance on the incomplete skeleton data with occlusion or noise. This paper addresses occluded and noise-robust skeleton-based action recognition and presents a novel Dual Inhibition Training strategy. Specifically, we propose Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN), which comprises of three parts: Input Skeleton Inhibition (ISI), Part-Aware Representation Learning (PARL) and Predicted Score Inhibition (PSI). The ISI and PSI are plug and play modules which could encourage the model to learn discriminative features from diversified body joints by effectively simulating key body part occlusions and random occlusions. The PARL module learns both the global and local representations from the whole body and body parts, respectively, and progressively fuses them during representation learning to enhance the model robustness under occlusions. Finally, we design different settings for occluded skeleton-based human action recognition to deep study this problem and better evaluate different approaches. Our approach achieves state-of-the-art results on different benchmarks and dramatically outperforms the recent skeleton-based action recognition approaches, especially under large-scale temporal occlusion.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision.*

KEYWORDS

Human Action Recognition, Skeleton Data, Noise-Robust

*Equal Contribution. †Corresponding authors: Jie Gui, Hongsong Wang.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612170

ACM Reference Format:
Zhenjie Chen*, Hongsong Wang* †, and Jie Gui†. 2023. Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612170

1 INTRODUCTION

Human action recognition possesses substantial application potential across various domains including video retrieval, smart home automation, entertainment, human-computer interaction, and security. In contrast to alternative modalities such as RGB and Flow, skeleton data offers a more concise structure and richer information. Skeleton data is robust to lighting and scene changes, and has low-dimensional representation which conserves computing resources. Owing to the advancements in depth sensors, skeleton data has become more accurate and accessible. These factors contribute to the widespread popularity of skeleton-based human action recognition and analysis [5, 6, 8, 15, 16, 19, 22, 31, 48–50, 53, 55].

Deep learning has made significant development, particularly in fields such as image recognition, natural language processing and speech recognition [12–14, 17, 30]. In skeleton-based human action recognition, the prevalent approach involves using deep learning models to simulate the spatial-temporal evolution of skeleton sequences [9, 58]. In earlier times, RNN or LSTM are utilized to model temporal contextual information within skeleton sequences, as they can handle dynamic dependencies in sequential data [26, 41, 51, 52]. Subsequently, CNN shows good learning capabilities in skeleton-based action recognition by converting the skeleton data into pseudo-images [10, 18, 27]. As skeleton data can be considered a type of graph data, representing skeleton data as a graph structure of edges and body joints can provide a better representation. Thus, GCN-based approaches [25, 37, 57, 58, 60] become popular and achieved remarkable success in recent years. However, these approaches are not effective in addressing challenges such as occlusion and multi-target interaction, which are commonplace in daily life. When certain crucial joints are obstructed or disturbed, the recognition ability of above models is significantly compromised.

With the fast development of techniques of skeleton-based action recognition, the robustness of models in occluded environments becomes an emerging obstacle. To tackle this issue, several studies skillfully design graph convolutional networks to enhance the

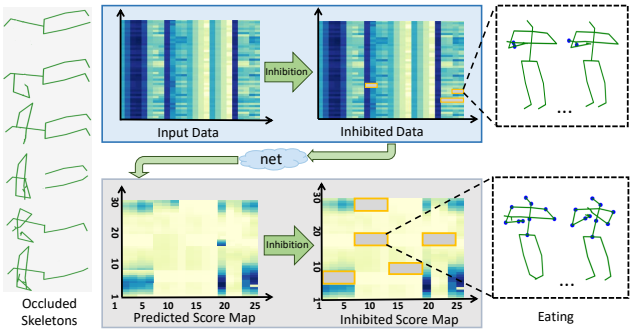

Figure 1: Dual Inhibition Training strategy. The area covered by the yellow rectangular block represents the inhibited area. On the right is the visualization of the skeleton corresponding to a certain inhibition area, where the blue dot indicates that the corresponding data or score is inhibited.

model's effectiveness on the noisy or incomplete skeletons, achieving favorable results [43, 45, 61]. However, these approaches are not general enough, and do not thoroughly analyze the model's robustness under different conditions.

Capturing incomplete or noisy skeletons is almost inevitable in real life scenarios. For example, individuals may be obscured by other entities, and the accuracy of human pose estimation can be compromised by environmental factors such as illumination fluctuations. Unfortunately, previous models fail to consider this aspect and primarily focus on mining discriminative body joints from the skeleton data. However, when these body joints are occluded, the recognition results experience significant deterioration. In light of this issue, we propose a Dual Inhibition Training strategy that seeks to simulate occlusion to compel the model to activate a larger number of body joints that may be potentially associated with actions and extract sufficient features from them. As shown in Figure 1, the net denotes to an extensive network. We inhibit some important body joints within the input data to simulate the situation where crucial parts are occluded. At the other end of the network, we partition the predicted score map into grids and randomly inhibit certain grids. This approach not only simulates random occlusion, but also increases the difficulty of prediction, forcing the model to learn as many features as possible from uninhibited joints.

Specifically, we apply the Dual Inhibition Training strategy to our proposed model: Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN), a novel method that improves recognition accuracy on occluded skeletons by simulating occlusion situations and fully extracting spatial information. The proposed method consists of three parts: Input Skeleton Inhibition (ISI), Part-Aware Representation Learning (PARL) and Predicted Score Inhibition (PSI). The ISI module simulates crucial part occlusion by inhibiting upstream activations on the input data. The PARL module carefully learns the potential association between joints and comprehensively extracts local information and global information from skeletons to enhance the model's robustness. The PSI module simulates random occlusion by inhibiting partial predicted score map, increasing the difficulty of model prediction and forcing the model to learn enough discriminative features from different joints. We also construct synthetic occlusion dataset that includes various occlusion scenarios and design several experimental settings to thoroughly test models from various perspectives. All results consistently validate the effectiveness and robustness of the proposed method on these settings with occluded datasets. Our model demonstrates superior performance compared to many competitive algorithms on several occluded benchmarks.

Our main contributions can be summarized as follows.

- We propose a novel Dual Inhibition Training strategy to simulate occlusion to increase the robustness of the model.
- We propose a Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN), a novel method that improves recognition accuracy in occlusion scenarios.
- We propose a Part-Aware Representation Learning (PARL) module to comprehensively extract spatial information.
- We construct synthetic occlusion datasets and design two experimental settings to thoroughly test models from various perspectives.

2 RELATED WORK

2.1 Skeleton-based Action Recognition

In recent years, skeleton-based action recognition has received extensive attention due to the advantages of skeleton data. For simplicity, we only review the GCN-based approaches which are the most similar to ours. Yan et al. [58] first introduce ST-GCN, which skillfully captures the spatial and temporal features by utilizing the attribute that skeleton data belongs to the graph structure. Li et al. [23] propose an Actional-Structural GCN which incorporates action links to capture specific potential dependencies among actions. Furthermore, it leverages structural links to enrich existing skeleton graphs, thus representing higher-order dependencies more comprehensively. Shi et al. [37] propose 2s-AGCN, which incorporates an adaptive graph convolution network and learns the topological structure of the graph dynamically. Subsequently, numerous studies have employed GCN as a fundamental framework [1, 36, 39, 62]. Zhang et al. [63] propose a context-aware graph convolution module, which aggregates information from adjacent joints with similar contexts. Peng et al. [33] apply neural architecture search to identify the optimal GCN architecture automatically. Liu et al. [28] introduce a novel disentangled multi-scale aggregation scheme and integrate it with G3D to create a robust feature extractor. Recently, Li et al. [24] introduce symbiotic GCNs to jointly learn the representation of both action recognition tasks and action prediction tasks. Chen et al. [3] develop a Channel-wise Topology Refinement Graph Convolution (CTR-GC) that refines the topology of skeleton data at the channel level. Song et al. [44] develop a novel framework that achieves state-of-the-art accuracy and high computational efficiency. Chi et al. [7] leverage the information bottleneck objective to guide the learning of action information and apply attention mechanism on graph convolution to capture context-sensitive information. Although these works have achieved satisfactory performances, they target at action recognition with complete skeletons and ignore the occluded or noisy skeleton data which are very common in realistic scenarios.

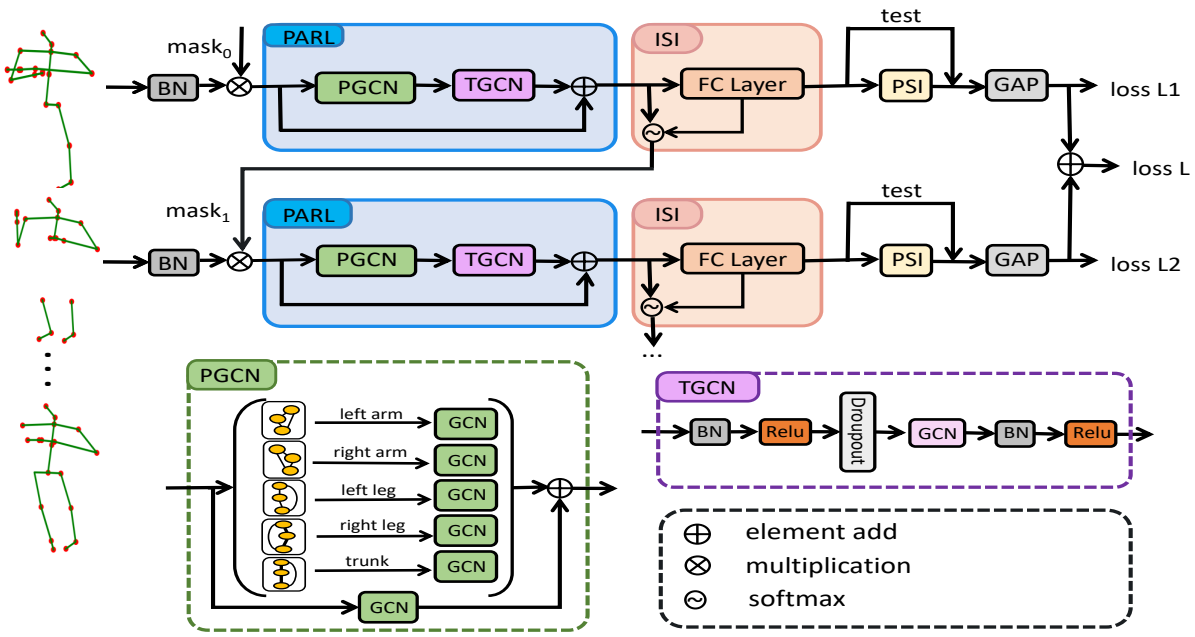

Figure 2: Framework of the proposed Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN). It is a multi-stream model and each stream contains three modules: PARL, ISI, and PSI. The PARL module is employed to extract detailed spatial features. ISI and PSI modules form the Dual Inhibition Training strategy, which simulates occlusion to enhance the model's robustness.

2.2 Occluded Human Action Recognition

In video-based human action recognition [21, 54], the significance of background and lighting conditions as environmental factors cannot be understated. To overcome this problem, Srinivasan et al. [46] extract the action information from the action vector generated during the video compression process, and the action vector is encoded using a histogram method. Ehatisham et al. [11] propose a feature-level fusion method for multi-modal data analysis. This approach combines the extracted features from different modalities into a unified representation, thereby enhancing the classifier's effectiveness. Moreover, Zuo et al. [65] incorporate the principles of fuzzy logic into conventional spatial-temporal features and introduce a novel histogram-based method. Broom et al. [2] investigate the essential role of time modeling in improving cross-domain robustness via comparative experiments, thus providing significant contributions to the field of video action recognition research. However, these studies are all focused on RGB videos and cannot be applied to skeletons.

Occluded skeletal-based action recognition has not received sufficient attention, and research in this domain is also scarce. In order to tackle this issue, Song et al. [43, 45] utilize class activation maps (CAM) [64] to quantify the degree of activation for each body joint across all GCN streams. To achieve extensive activation, they further apply a masking technique for active joints. Shi et al. [38] introduce a novel multi-stream model that seeks to improve recognition accuracy on occlusion datasets. It effectively handles various occlusion scenarios by utilizing different streams, and ultimately fuses the features of these streams to achieve superior results. Song et al. [42] regard this problem as skeleton noise-adaptation, and build different models that extract noise-robust features from paired or unpaired noisy skeletons. Peng et al. [32] consider the task of one-shot skeleton-based action recognition, and present a transformer-based model which takes image-like representations of joints, bones, and velocities. Although the aforementioned approaches have somewhat enhanced the model's robustness, they still exhibit certain limitations in terms of intricate design and a dearth of universality.

3 METHOD

Human body can be partly or completely occluded in realistic scenarios, resulting in reduced recognition efficiency. However, the majority of current skeleton-based action recognition methods prioritize boosting performances on the non-occluded data, while neglecting the model's capacity to sustain robustness against occlusion or noise.

To address noise-robust skeleton-based action recognition, we propose Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN). On the one hand, we simulate occlusion by employing Dual Inhibition Training strategy which is divided into Input Skeleton Inhibition (ISI) module and Predicted Score Inhibition (PSI) module. On the other hand, we extract detailed spatial features by using Part-Aware Representation Learning (PARL). It should be noted that during the test phase, we do not inhibit the predicted score map, which means that all information regions detected during the training stage contribute to the final confidence

score. Figure 2 shows the overall architecture of our model. Details are described in the following sections.

3.1 Input Skeleton Inhibition

To enforce the network to learn robust representations from different joints, the ISI module straightly simulates the occluded or noisy skeleton data during training. Suppose a human skeleton data consists of J body joints and T time steps. We simply mask random joints of the input skeleton at different time steps with the formulation as

$$\tilde{\mathbf{X}} = \mathbf{X} \otimes \mathbf{mask}, \tag{1}$$

where $\mathbf{X}$ is the original skeletons which is a $T \times J$ matrix, **mask** is the float mask matrix with the same dimensionality, and $\otimes$ denotes element multiplication.

Inspired by Grad-CAM [34, 43], we formulate the mask based on class activation maps which highlight discriminative human joints. The ISI module solely consists of a fully connected layer (FC) and a softmax activation function. Suppose $\mathbf{F(t, j)}$ is the feature vector before the FC layer for the j-th joint and the t-th time step, and $\mathbf{W_c}$ is the weight for the c-th action class of the FC layer. The activation value for the j-th joint and the t-th time step is

$$S(t, j) = \sum_c \mathbf{W}_c \mathbf{F(t, j)}, \tag{2}$$

where $S(t, j)$ is an element of the activation matrix $\mathbf{S}$.

Given the $T \times J$ activation matrix $\mathbf{S}$, a softmax activation function is applied to normalize the values between 0 and 1. We obtain the normalization mask matrix M by

$$\mathrm{M} = 1 - \mathrm{softmax}(\mathbf{S}), \tag{3}$$

where M and $\mathbf{S}$ have the same dimensionality, and the value of M represents the degree of inhibiting. It should be noted that when the activation value is very high, the value of mask matrix is close to zero, indicating that the input data at that position will be almost completely masked.

To ensure the model's robust performance even in the event of critical information loss, we only input body joints that have not been activated by the upper stream, forcing the model to learn enough discriminative features from unmasked joints. The final s+1 stream float mask matrix is calculated by

$$\mathrm{mask_s} = (\prod_{\mathrm{i}=1}^{\mathrm{s}-1} \mathrm{mask_i}) \otimes \mathrm{M_{s-1}}, \tag{4}$$

where $\prod$ and $\otimes$ both denote element multiplication.

By the way, to achieve maximum consistency in input data distribution, we employ a random generation method to produce $\mathrm{mask_0}$ for the first stream. Moreover, to avoid interference with joint activation in the subsequent stream, we exclude $\mathrm{mask_0}$ from the computation of the float mask matrix in the downstream processing.

3.2 Part-Aware Representation Learning

It is evident that the majority human actions do not involve the entirety of body, but rather are the results of focused actions performed by particular body parts. For example, brushing teeth, shaking hands and writing can be determined by the movement of joints of the hand. This fact motivates us to highlight the localized characteristics of human actions.

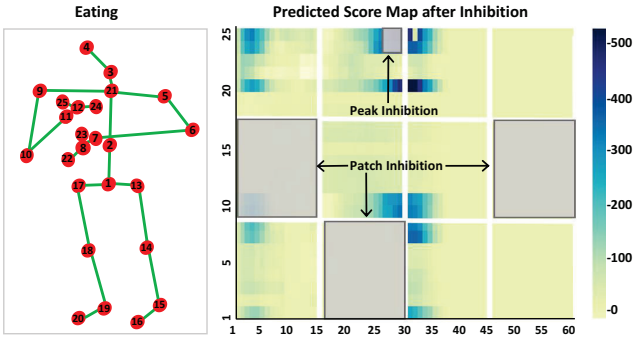

Figure 3: Illustration of Predicted Score Inhibition (PSI). The horizontal axis depicts a temporal sequence, while the longitudinal axis portrays 25 skeletal joints. The positions of these joints in the human body are shown on the left.

In order to capture these local features with greater precision, previous work (such as ST-GCN [58]) has pre-defined graph connections or edges of human joints and partitions them based on the distance between adjacent joints and root joints to simulate local differential properties.

To further leverage the local characteristics of actions and learn more detailed features, we propose Part-Aware Representation Learning (PARL). The PARL Block is composed of PGCN, TGCN, and a residual connection. The TGCN denotes temporal GCN, which is the same as in ST-GCN. PGCN is designed to capture detailed spatial features. Specifically, in PGCN, we divide the input data into multiple parts based on the joint semantic information, and construct a corresponding topology graph by treating the joints of each part as fully connected. Each part $\mathbf{X}_k$ enters a different GCN to extract local information. Moreover, to avoid missing global information, similar to STGCN, we pre-define a graph based on joint connections and feed the entire human joint data $\mathbf{X}_0$ into GCN to obtain the global relationship features $\mathbf{H}_0$. The outputs of each GCN $\mathbf{H}_k$ is calculated by

$$\mathbf{H}_k = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}}_k \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X}_k \mathbf{W}_k), \tag{5}$$

where k denotes the number of the part. $\mathbf{X}_k$ represents partial human body part, wherein $\mathbf{X}_0$ represents the entire body. $\hat{\mathbf{D}}$ denotes the degree matrix. $\hat{\mathbf{A}}_k = \mathbf{A}_k + \mathbf{I}$, $\mathbf{A}_k$ represents the adjacency matrix. In particular, $\mathbf{A}_k$ for all parts are 1-matrix. $\mathbf{W}_k$ is the weight matrix learned by network. Then we merge the output of GCN with the following formula

$$\mathbf{H} = \alpha \sum_{k=1}^{m} \mathbf{H}_k + (1 - \alpha)\mathbf{H}_0, \tag{6}$$

where α is a proportional parameter, m denotes the number of parts, and $\mathbf{H}$ is the output of PGCN. In this way, the network can comprehensively learn both the global and local features of skeletons.

In addition, PARL module stacked, learning potential discriminative features in different layers from shallow to deep in the network. These enable the model to extract sufficiently effective information from other joints, even when some body joints are occluded, thereby providing a significant advantage.

3.3 Predicted Score Inhibition

As we all know, occlusion can occur randomly with uncertain timing and location. In order to improve the model's robustness against such unpredictable occlusion, we devise a Predicted Score Inhibition (PSI) which randomly inhibits the predicted scores map in both spatial and temporal dimensions of skeleton data. The proposed PSI takes inspiration from the diversification block of fine-grained image recognition [47] which aims to learn subtle differences between similar object classes. It is divided into two components: peak inhibition and patch inhibition. The illustration of PSI module is shown in Figure 3.

Peak Inhibition. Suppose $\mathbf{P} \in \mathbb{R}^{C\times T\times J}$ is the output predicted score map of the FC layer, where C denotes the number of action classes, T and J represent the length of the skeleton sequence and the number of body joints, respectively. To further inhibit the key joints, we detect peak value location of $\mathbf{P}$ because it is the most discriminative regions for the classifier, and apply a probability of p_{peak} to inhibit it. Define $\mathbf{S}' \in \mathbb{R}^{T\times J}$ as the binary peak inhibition matrix of $\mathbf{P}$, where 1 indicates the peak location will be inhibited, while 0 means no inhibition will take place. The $\mathbf{S}'$ is calculated by

$$S'(t,j) = \begin{cases} r, \text{if } P(t,j) = \max(\mathbf{P}), r \sim \text{Bernoulli}(p_{peak}), \\ 0, \text{otherwise}, \end{cases} \tag{7}$$

where $P(t,j)$ and $S'(t,j)$ are elements of the predicted score map $\mathbf{P}$ and peak inhibition matrix $\mathbf{S}'$, respectively. The $\max(\cdot)$ denotes the maximum function, and r is a Bernoulli random variable that equals 1 with the probability of p_{peak}.

Patch Inhibition. For other discriminative regions, taking into account cases where certain parts of the body are occluded across multiple consecutive frames, we design patch inhibition. Specifically, for the given predicted score map $\mathbf{P}$, we divide it into grids along the spatial and temporal dimensions. The $\mathbf{patch}^{[l,m]} \in \mathbb{R}^{t'\times j'}$ is the resulting patch, where $l \in [1...\frac{T}{t'}]$, $m \in [1...\frac{J}{j'}]$, t' is the length of patch skeleton sequence, and j' is the number of patch body joints. We inhibit each $\mathbf{patch}^{[l,m]}$ with probability p_{patch}. Suppose $\mathbf{S}'' \in \mathbb{R}^{T\times J}$ is the binary random patch inhibition matrix. If a patch is inhibited, the patch inhibition matrix corresponding to this patch is set to $\mathbf{1}$, otherwise, it is set to $\mathbf{0}$. Thus the $\mathbf{S}''$ is obtained as

$$\mathbf{S}'' = \{\mathbf{patch}^{[l,m]} \in [\mathbf{0},\mathbf{1}] : l = 1...\frac{T}{t'}, m = 1...\frac{J}{j'}\}. \tag{8}$$

The overall inhibition matrix $\mathbf{S} \in \mathbb{R}^{T\times J}$ is the combination of both peak inhibition matrix and patch inhibition matrix. It is obtained by merging $\mathbf{S}'$ and $\mathbf{S}''$ as

$$\mathbf{S} = 1 - \beta \cdot (\mathbf{S}'|\mathbf{S}''), \tag{9}$$

where β is the inhibition factor and $|$ denotes logic or operator. Finally, the inhibited score map $\tilde{\mathbf{P}}$ is generated as

$$\tilde{\mathbf{P}} = \mathbf{S} \otimes \mathbf{P}, \tag{10}$$

where $\otimes$ represents element level multiplication.

By utilizing the methods of the peak inhibition and patch inhibition, we elevate the learning challenges faced by the model and encourage the acquisition of discriminative features from uninhibited joints, thereby improving the robustness of the model. Furthermore, it should be noted that the PSI module is employed only in the training phase. During the test phase, the complete predicted score map $\mathbf{P}$ is directly transferred to the global average pooling without any inhibition in any region. In this way, all information regions detected during the training stage contribute to the final confidence score.

4 EXPERIMENTS

In this section, we design two different settings for occluded skeleton-based human action recognition. We compare our model with other state-of-the-art methods and perform a detailed ablation study to confirm the efficacy of the proposed components.

4.1 Experimental Settings

To thoroughly test the occluded skeleton-based action recognition from various perspectives, we establish two experimental settings: Robustness to Occlusions and Recognition on Synthetic Data.

Robustness to Occlusions. In accordance with the experimental setting [45], we train models on the training set of NTU RGB+D 60 dataset [35] and set up two types of occluded test: spatial occlusion test and temporal occlusion test. For the former, we evaluate models using skeletons without joints of left arm, right arm, two hands, two legs, and trunk, respectively. For the latter, we evaluate models with skeletons in which blocks of frames are randomly occluded. The number of occluded frames are 10, 20, 30, 40, and 50, respectively. By learning on complete data and testing on incomplete data, we can validate the robustness of the model.

NTU RGB+D 60 dataset [35] contains 60 distinct categories of actions, amounting to 56,880 samples. The movements are performed by 40 volunteers aged between 10 and 35 years. The dataset is collected by Microsoft Kinect v2 sensor, along with three cameras positioned at different angles. The collected data comprises depth video, human skeletons, RGB video and infrared sequence. A human skeleton is represented by 25 joints with 3D coordinates.

Recognition on Synthetic Data. The unpredictability of occlusion poses a significant challenge to human action recognition, necessitating the urgent need for datasets that support occlusion experiments. In this experimental setting, we create a synthetic occlusion dataset which contains multiple occlusion scenarios to simulate realistic occlusions.

We randomly select 50% of the samples from the cross-subject benchmark of NTU RGB+D 60 [35] for occlusion. The occlusion scenes are evenly divided into five categories: left arm occlusion, right arm occlusion , two hands occlusion, lower limb occlusion, and trunk occlusion. Furthermore, we retain the remaining complete data since not all cases involve occlusion. Therefore, the synthetic occlusion dataset is composed of an equal proportion of complete and occlusion data. In addition, five occlusion scenes are evenly distributed within the occlusion data. We train models on the training set of the synthetic occlusion dataset and set two test sets: spatial occlusion test, temporal occlusion test. The test sets are the same as those of the experimental setting of Robustness to Occlusion.

Models are trained on the incomplete occlusion data and also tested on the incomplete occlusion data. In this way, we can ascertain the ability of the model to acquire adequate discriminative features from occluded skeleton data and the capacity to solve action recognition problems even in obscured conditions. For fair

Table 1: Performance comparison of spatial occlusion on the NTU RGB+D 60 dataset in terms of accuracy (%).

Spatial	Occluded Part						
Occlusion	None	Left Arm	Right Arm	Two Hands	Two Legs	Trunk	Mean
ST-GCN [58]	80.7	71.4	60.5	62.6	77.4	50.2	64.4
SR-TSL [40]	84.8	70.6	54.3	48.6	74.3	56.2	60.8
2s-AGCN [37]	88.5	72.4	55.8	**82.1**	74.1	71.9	71.3
1s RA-GCN [43]	85.8	69.9	54.0	66.8	82.4	64.9	67.6
2s RA-GCN [43]	86.7	75.9	**62.1**	69.2	83.3	72.8	72.7
3s RA-GCN [43]	87.3	74.5	59.4	74.2	83.2	72.3	72.2
STIGCN [20]	88.8	12.7	11.5	18.3	45.5	20.9	21.8
MS-G3D[29]	87.3	31.3	23.8	17.1	78.3	61.6	42.4
CTR-GCN[4]	87.5	13.0	12.5	12.7	21.0	36.3	19.1
TCA-GCN[56]	**90.2**	75.4	53.4	70.8	75.2	**78.6**	70.7
HD-GCN[59]	86.8	67.1	55.7	56.7	74.8	61.3	63.1
1s PDGCN (ours)	85.7	73.4	60.4	65.9	83.0	71.2	70.8
2s PDGCN (ours)	87.4	**76.4**	62.0	74.4	84.8	70.4	73.6
3s PDGCN (ours)	87.5	76.0	62.0	75.4	**85.0**	73.0	**74.3**

Mean : the average accuracy of occluding the left arm, right arm, two hands, two legs, and the trunk.

Table 2: Performance comparison of temporal occlusion on the NTU RGB+D 60 dataset in terms of accuracy (%).

Temporal	Number of Occluded Frames						
Occlusion	0	10	20	30	40	50	Mean
ST-GCN [58]	80.7	69.3	57.0	44.5	34.5	24.0	45.9
SR-TSL [40]	84.8	70.9	62.6	48.8	41.3	28.8	50.5
2s-AGCN [37]	88.5	74.8	60.8	49.7	38.2	28.0	50.3
1s RA-GCN [43]	85.8	81.6	72.9	61.6	47.9	34.0	59.6
2s RA-GCN [43]	86.7	83.0	76.4	65.6	53.1	39.5	63.5
3s RA-GCN [43]	87.3	83.9	76.4	66.3	53.2	38.5	63.7
STIGCN [20]	88.8	70.4	51.0	38.7	23.8	8.0	38.4
MS-G3D[29]	87.3	77.6	65.7	54.3	41.9	30.1	53.9
CTR-GCN[4]	87.5	72.4	54.1	35.6	22.4	11.5	39.2
TCA-GCN[56]	**90.2**	**84.4**	74.6	58.1	42.3	25.6	57.0
HD-GCN[59]	86.8	57.0	29.5	18.5	11.2	7.04	24.7
1s PDGCN (ours)	85.7	81.9	75.4	66.4	54.9	40.0	63.7
2s PDGCN (ours)	87.4	83.8	**76.7**	**66.8**	**55.1**	**40.6**	**64.6**
3s PDGCN (ours)	87.5	83.9	76.6	66.7	53.9	40.0	64.2

Mean : the average accuracy of occlusion for 10, 20, 30, 40,and 50 frames.

comparisons, we employ the same data preprocessing for all models to eliminate the impact of data preprocessing.

4.2 Implementation Details

Network Setting. To normalize the input data, a data BN layer is employed at the front end of the network. In the PGCN module of the PARL block, the human body is subdivided into five non-overlapping parts, i.e., left arm, right arm, left leg, right leg, trunk. We view the joints of each part as interconnected to learn potential associations. The proportional parameter α is set to 0.1. In the ISI, the number of neurons in the FC layer is established as the number of categories. In the PSI, the inhibition probability and patch length and width are designated as hyperparameters, which can be adjusted according to the characteristics of different datasets. In our experiment, p_{peak} is set to 0.9, p_{patch} is set to 0.2, and β is set to 0.1. The patch length t' and width j' are set to 5 and 10, respectively. Finally, prediction results for each stream are obtained by applying an average global pooling to the predicted score map along the temporal and spatial dimensions.

Table 3: Performance comparison of spatial occlusion on synthetic occlusion dataset in terms of accuracy (%).

Spatial	Occluded Part						
Occlusion	None	Left Arm	Right Arm	Two Hands	Two Legs	Trunk	Mean
ST-GCN [58]	83.8	76.9	67.9	83.5	80.5	80.6	77.9
2s-AGCN [37]	84.5	75.9	66.9	83.0	81.8	81.1	77.7
1s RA-GCN [43]	83.8	76.3	63.5	79.4	81.6	80.0	76.1
2s RA-GCN [43]	84.8	77.1	70.5	82.6	83.2	81.7	79.0
3s RA-GCN [43]	85.6	79.1	**71.3**	85.0	83.5	82.5	80.3
1s PDGCN (ours)	83.4	76.0	68.2	82.6	82.5	79.7	77.8
2s PDGCN (ours)	85.9	79.3	70.4	**85.3**	83.8	82.3	80.2
3s PDGCN (ours)	**86.0**	**79.4**	70.3	85.1	**84.1**	**83.1**	**80.4**

Mean : the average accuracy of occluding the left arm, right arm, two hands, two legs, and the trunk.

Table 4: Performance comparison of temporal occlusion on synthetic occlusion dataset in terms of accuracy (%).

Temporal	Number of Occluded Frames						
Occlusion	0	10	20	30	40	50	Mean
ST-GCN [58]	83.8	78.7	67.7	54.3	41.6	29.9	54.4
2s-AGCN [37]	84.5	77.1	66.1	54.8	43.5	30.9	54.5
1s RA-GCN [43]	83.8	78.0	67.8	56.9	45.4	33.0	56.2
2s RA-GCN [43]	84.8	80.8	72.1	61.6	49.1	35.2	59.8
3s RA-GCN [43]	85.6	81.5	72.9	61.6	49.3	36.2	60.3
1s PDGCN (ours)	83.4	78.4	70.1	60.4	49.4	38.2	59.3
2s PDGCN (ours)	85.9	81.8	**74.1**	**64.5**	**52.8**	**39.4**	**62.5**
3s PDGCN (ours)	**86.0**	**82.1**	73.5	62.8	51.4	38.5	61.7

Mean : the average accuracy of occlusion for 10, 20, 30, 40, and 50 frames

Training. All experiments are conducted on PyTorch deep learning platform with one GPU. The optimization strategy is stochastic gradient descent (SGD) with a momentum of 0.9. The batch size is 16. The network is trained with the cross-entropy loss. On NTU RGB+D 60 dataset and synthetic occluded dataset, pretraining parameters provided by RA-GCN [43] are utilized. The initial learning rate is 0.1, and on the 10-th and 30-th epochs, it is reduced by a factor of 10. The maximum number of training epochs is 60.

Data Processing. For the NTU RGB+D 60 dataset, each sample has the maximum number of two human skeletons. If a sample only has one human skeleton, we add a padding of 0 to the second skeleton. Every sample consists of precisely 300 frames. If a sample contains fewer than 300 frames, we replicate it until it reaches 300 frames. Furthermore, we apply the same preprocessing module as RA-GCN [43] to transform the input data, so that the converted data encompasses position coordinate information, relative coordinate information, and relative motion information.

4.3 Experimental Results

Results under the above experimental settings are presented and state-of-the-art approaches are compared to demonstrate the robustness of our model on incomplete skeleton data.

Robustness to Occlusions. In this setting, we train the model on complete skeleton data and test it on the incomplete data. We evaluate the model by occluding these different parts on the test set, and results are shown in Table 1. It is observed that, in general, our model achieves excellent performance when tested on the dataset with body parts occlusion. Compared with 1s RA-GCN [43], the proposed 1s PDGCN has a considerable improvement. Furthermore,

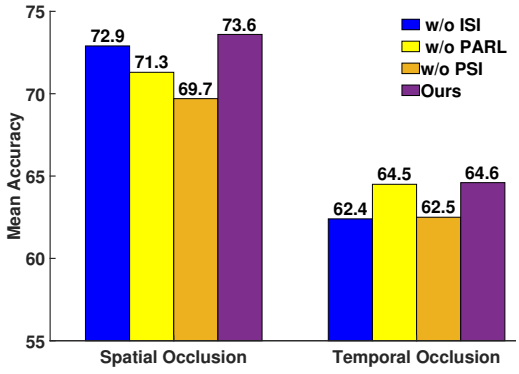

Figure 4: Ablation studies of the PDGCN on cross-subject benchmark of the NTU RGB+D 60 dataset.

as the number of streams increases, the average accuracy of our model also improves. It can be concluded that the PDGCN is robust to part occlusion and the robustness would improve with the increase of model streams.

To simulate temporal occlusion, we randomly occlude a subsequence within the first 100 frames on the test data and set the occluded window size to be 10, 20, 30, 40, and 50, respectively. As illustrated in Table 2, our 2s PDGCN model outperforms other methods in terms of accuracy on occluded frames with different lengths. Furthermore, the performance gap between our model and RA-GCN model increases as the number of occluded frames increases, indicating significant advantages of our model in scenarios with a substantial amount of temporal occlusion.

It should noted that, as most state-of-the-art skeleton-based action recognition models such as STIGCN [20], TCA-GCN[56] and HD-GCN[59] are not specifically designed to tackle occlusions, they have poor performance on the occluded benchmarks.

Recognition on Synthetic Data. To further enhance the occlusion robustness and ensure more fair comparisons, we add the preprocessing module for all models and train the model on synthetic occlusion dataset that we created. Results of spatial occlusion test are summarized in Table 3. Our model achieves the highest accuracy compared to other models. Compared to the results in Table 1, the model's accuracy has increased by an average of seven points, proving effectiveness of the proposed training strategy.

Table 4 shows results of temporal occlusion test. Our model also achieves state-of-the-art performance. In comparison to results in Table 2, the recognition accuracy of the model decreases slightly. It is comprehensible because our synthetic dataset does not contain samples with temporal occlusion, and the spatial occlusion weakens the learning of temporal features. Compared to the improved spatial occlusion accuracy, this slight reduction is acceptable.

4.4 Ablation Study

Without loss of generality, we evaluate the effectiveness of components proposed by PDGCN under both the spatial and temporal occlusions of the setting of robustness to occlusions. The ablated components are Input Skeleton Inhibition (ISI), Part-Aware Representation Learning (PARL) and Predicted Score Inhibition (PSI).

Table 5: Comparisons of efficiency and parameters.

Method	FLOPs	Parameters
ST-GCN (base model) [58]	16.3 G	3.1 M
1s-PDGCN (w/ PARL)	10.9 G	2.0 M
2s-PDGCN (w/ PARL)	21.9 G	4.1 M
1s-PDGCN (ours)	12.3 G	3.1 M
2s-PDGCN (ours)	24.6 G	6.3 M

The results are shown in Figure 4. By eliminating the ISI module from PDGCN, we observe a decrease of 0.7% in average accuracy of spatial occlusion and a 2.2% decline in average accuracy of temporal occlusion. Without the PARL module, we notice a 2.3% reduction in average accuracy of spatial occlusion, while the average accuracy for temporal occlusion only experiences a slight decrease. It indicates that the design of the PARL module contributes more effectively to the model's spatial feature learning. The PSI module, as an essential component of Dual-inhibition Training strategy, significantly influences the model's performance. Removing the PSI module results in a 3.9% decrease in average accuracy of spatial occlusion and a 2.1% decline in average accuracy of temporal occlusion. By combining these three modules, our model achieves the highest recognition accuracy.

The efficiency of the proposed method is analyzed in Table 5. Since our model is built on the ST-GCN [58], we only compare computational cost and model parameters with the ST-GCN. The ISI and PSI modules do not add additional model parameters, while only a small number of parameters are added by the PARL module. The proposed training strategy almost not add additional model parameters, and the computational complexity of our model is the same as the base model.

4.5 Visualizations

We visualize the functions of the PARL module, and present visualizations of occluded samples of human skeletons along with recognition results for different models.

Visualization of Joint Activation. For analyzing the role of PARL module in our approach, we visualize the activation status of each joint, and compare it with the model without the PARL. As depicted in Figure 5, we visualize the joint activation patterns in each stream for three typical actions: brushing teeth, eating, and saluting. Large red circles represent the top-5 activated joints of each body, while small red circles represent the remaining joints. We observe that compared to the model without PARL, the model employing PARL promptly focuses on the crucial joints closely associated with the actions. For example, for the brushing teeth action, it is obvious that hand movements are more important than foot movements. The model with PARL first focuses on hand joints, whereas the model without PARL only focuses on feet in the first stream. We also observe significant differences between the activated joints in the first and second streams. It can be explained the ISI module inhibits the input data corresponding to the activation position in the previous streams, forcing the model to activate other potentially valuable joints, thereby enhancing the model's occlusion robustness.

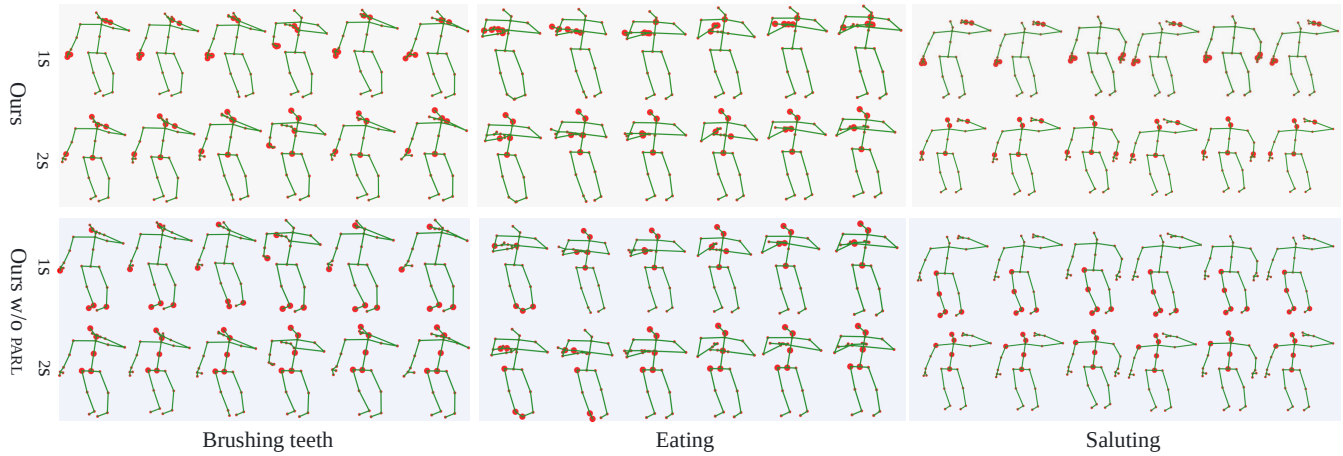

Figure 5: Visualization of body joint activation in different streams for the proposed method with or without the PARL module. 1S and 2S denote the first and second stream, respectively.

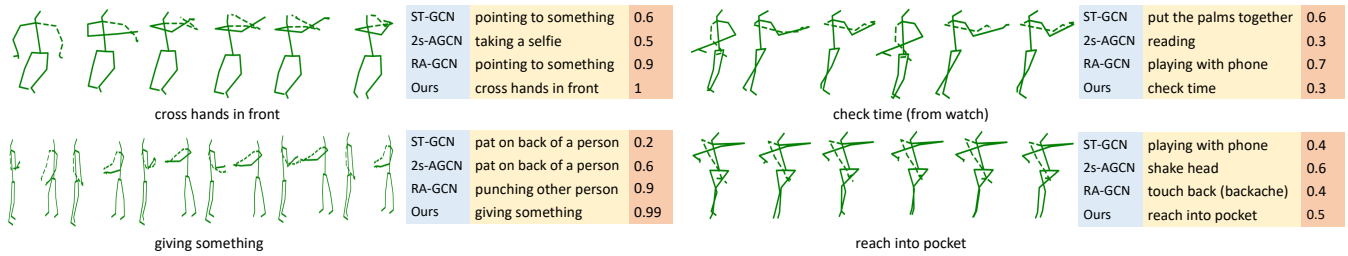

Figure 6: Visualization of occluded samples and corresponding classification results across different models. The dashed line indicates the occluded region. For each sample, predicted categories as well as confidence scores are displayed on the left.

Visualization of Recognition Results of Occluded Skeletons. We compare the recognition results of our approach with those of the state-of-the-art approaches for occluded skeleton samples in Figure 6. The results show that our model performs well on some challenging samples compared to the other models. For instance, regarding the two samples situated on the left, our model successfully classifies them with high confidence, whereas other models inaccurately assign them to different categories with relatively high confidence. For the two samples on the right, despite a comparatively lower confidence score, our model accurately classifies them, demonstrating the significant effectiveness of our model in handling occluded data. In addition, the model's classification outcomes depicted in the figure also highlight various actions that are prone to confusion when the left arm is obstructed. Even for human observers, discerning these actions becomes arduous when presented as incomplete skeletal sequences. Therefore, it becomes evident that visually indistinguishable actions pose a huge challenge.

5 CONCLUSION

In this paper, we address occluded skeleton-based human action recognition and present an effective training method called Dual Inhibition Training. We propose a novel Part-aware and Dual-inhibition Graph Convolutional Network (PDGCN) which consists of Input Skeleton Inhibition, Part-Aware Representation Learning and Predicted Score Inhibition. To ensure a comprehensive evaluation, we establish several experimental settings. The proposed PDGCN consistently achieves state-of-the-art performance on these settings under different occlusions. Ablation studies demonstrate the effectiveness of each component of the proposed method and indicate that simulating occlusion in both temporal and spatial dimensions could enhance the robustness. Visualizations show that our approach could accurately recognize some heavily occluded samples where the other approaches fail. In the future, we will explore deeper into the problem and further improve distinctions between similar categories. We believe our work will inspire further research in this area.

ACKNOWLEDGMENTS

This work is jointly supported by the NSFC Grant No. 62172090, the CAAI-Huawei MindSpore Open Fund, the Fundamental Research Funds for the Central Universities No. 2242023K30034, the Start-up Research Fund of Southeast University under Grant RF1028623097 and Grant RF1028623063. This work is also supported by the Big Data Computing Center of Southeast University.

REFERENCES

[1] Cunling Bian, Wei Feng, and Song Wang. 2022. Self-supervised representation learning for skeleton-based group activity recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5990–5998.

[2] Sofia Broomé, Ernest Pokropek, Boyu Li, and Hedvig Kjellström. 2023. Recur, Attend or Convolve? On Whether Temporal Modeling Matters for Cross-Domain Robustness in Action Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 4199–4209.

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

[4] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. 2021. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. In *2021 IEEE/CVF International Conference on Computer Vision*. 13339–13348. https://doi.org/10.1109/ICCV48922.2021.01311

[5] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. 2022. Hierarchically self-supervised transformer for human skeleton representation learning. In *European Conference on Computer Vision*. Springer, 185–202.

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

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

[8] Yunfeng Diao, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and He Wang. 2021. BASAR: Black-box attack on skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7597–7607.

[9] Yong Du, Yun Fu, and Liang Wang. 2016. Representation learning of temporal dynamics for skeleton-based action recognition. *IEEE Transactions on Image Processing* 25, 7 (2016), 3010–3022.

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

[11] Muhammad Ehatisham-Ul-Haq, Ali Javed, Muhammad Awais Azam, Hafiz MA Malik, Aun Irtaza, Ik Hyun Lee, and Muhammad Tariq Mahmood. 2019. Robust human activity recognition using multimodal feature-level fusion. *IEEE Access* 7 (2019), 60736–60751.

[12] Alex Graves and Navdeep Jaitly. 2014. Towards End-to-End Speech Recognition with Recurrent Neural Networks. In *Proceedings of the International Conference on International Conference on Machine Learning* (Beijing, China) *(ICML'14)*. JMLR.org, 1764–1772.

[13] Jie Gui, Xiaofeng Cong, Yuan Cao, Wenqi Ren, Jun Zhang, Jing Zhang, Jiuxin Cao, and Dacheng Tao. 2023. A Comprehensive Survey and Taxonomy on Single Image Dehazing Based on Deep Learning. *ACM Comput. Surv.* 55, 13s, Article 279 (jul 2023), 37 pages. https://doi.org/10.1145/3576918

[14] Jie Gui, Dacheng Tao, Zhenan Sun, Yong Luo, Xinge You, and Yuan Yan Tang. 2014. Group Sparse Multiview Patch Alignment Framework With View Consistency for Image Classification. *Image Processing* (2014).

[15] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. 2022. Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 762–770.

[16] Fei Han, Brian Reily, William Hoff, and Hao Zhang. 2017. Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding* 158 (2017), 85–105.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (2012), 82–97.

[18] Yonghong Hou, Zhaoyang Li, Pichao Wang, and Wanqing Li. 2016. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 3 (2016), 807–811.

[19] Yilei Hua, Wenhan Wu, Ce Zheng, Aidong Lu, Mengyuan Liu, Chen Chen, and Shiqian Wu. 2023. Part Aware Contrastive Learning for Self-Supervised Action Recognition. *arXiv preprint arXiv:2305.00666* (2023).

[20] Zhen Huang, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 2020. Spatio-temporal inception graph convolutional networks for skeleton-based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2122–2130.

[21] Yu Kong and Yun Fu. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision* 130, 5 (2022), 1366–1401.

[22] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4741–4750.

[23] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3595–3603.

[24] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2021), 3316–3333.

[25] Tingtian Li, Zixun Sun, and Xiao Chen. 2020. Group-Skeleton-Based Human Action Recognition in Complex Events. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4703–4707.

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

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

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

[29] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. 2020. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[30] Daniel W. Otter, Julian R. Medina, and Jugal K. Kalita. 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Transactions on Neural Networks and Learning Systems* 2 (2021), 32.

[31] Yunsheng Pang, Qiuhong Ke, Hossein Rahmani, James Bailey, and Jun Liu. 2022. Igformer: Interaction graph transformer for skeleton-based human interaction recognition. In *European Conference on Computer Vision*. Springer, 605–622.

[32] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. 2023. Delving deep into one-shot skeleton-based action recognition with diverse occlusions. *IEEE Transactions on Multimedia* (2023).

[33] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. 2020. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2669–2676.

[34] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.

[35] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1010–1019.

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

[37] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12026–12035.

[38] Wuzhen Shi, Dan Li, Yang Wen, and Wu Yang. 2023. Occlusion-Aware Graph Neural Networks for Skeleton Action Recognition. *IEEE Transactions on Industrial Informatics* (2023).

[39] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. 2019. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1227–1236.

[40] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Skeleton-based action recognition with spatial reasoning and temporal stack learning. In *Proceedings of the European conference on computer vision (ECCV)*. 103–118.

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

[42] Sijie Song, Jiaying Liu, Lilang Lin, and Zongming Guo. 2021. Learning to Recognize Human Actions From Noisy Skeleton Data Via Noise Adaptation. *IEEE Transactions on Multimedia* 24 (2021), 1152–1163.

[43] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2020. Richly activated graph convolutional network for robust skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 5 (2020), 1915–1925.

[44] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2022. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 1474–1488.

[45] Yi-Fan Song, Zhang Zhang, and Liang Wang. 2019. Richly activated graph convolutional network for action recognition with incomplete skeletons. In *IEEE International Conference on Image Processing*. IEEE, 1–5.

[46] Vignesh Srinivasan, Serhan Gul, Sebastian Bosse, Jan Timo Meyer, Thomas Schierl, Cornelius Hellge, and Wojciech Samek. 2016. On the robustness of action recognition methods in compressed and pixel domain. In *European Workshop on Visual Information Processing (EUVIP)*. IEEE, 1–6.

[47] Guolei Sun, Hisham Cholakkal, Salman Khan, Fahad Khan, and Ling Shao. 2020. Fine-grained recognition: Accounting for subtle differences between similar classes. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12047–12054.

[48] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. 2022. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[49] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1655–1663.

[50] Hongsong Wang, Jian Dong, Bin Cheng, and Jiashi Feng. 2021. PVRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. *IEEE Transactions on Image Processing* 30 (2021), 6096–6106.

[51] Hongsong Wang and Liang Wang. 2017. Modeling Temporal Dynamics and Spatial Configurations of Actions Using Two-Stream Recurrent Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[52] Hongsong Wang and Liang Wang. 2018. Beyond Joints: Learning Representations From Primitive Geometries for Skeleton-Based Action Recognition and Detection. *IEEE Transactions on Image Processing* 27, 9 (2018), 4382–4394.

[53] Hongsong Wang and Liang Wang. 2018. Learning content and style: Joint action recognition and person identification from human skeletons. *Pattern Recognition* 81 (2018), 23–35.

[54] Lei Wang, Du Q Huynh, and Piotr Koniusz. 2019. A comparative review of recent kinect-based action recognition algorithms. *IEEE Transactions on Image Processing* 29 (2019), 15–28.

[55] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. 2018. RGB-D-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding* 171 (2018), 118–139.

[56] Shengqin Wang, Yongji Zhang, Minghao Zhao, Hong Qi, Kai Wang, Fenglin Wei, and Yu Jiang. 2022. Skeleton-based Action Recognition via Temporal-Channel Aggregation. arXiv:2205.15936 [cs.CV]

[57] Xuanhan Wang, Yan Dai, Lianli Gao, and Jingkuan Song. 2022. Skeleton-based action recognition via adaptive cross-form learning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1670–1678.

[58] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[59] Zhi Yang, Kang Li, Haitao Gan, Zhongwei Huang, and Ming Shi. 2023. HD-GCN: A Hybrid Diffusion Graph Convolutional Network. *arXiv preprint arXiv:2303.17966* (2023).

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

[61] Yongsang Yoon, Jongmin Yu, and Moongu Jeon. 2022. Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition. *Applied Intelligence* (2022), 1–15.

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

[63] Xikun Zhang, Chang Xu, and Dacheng Tao. 2020. Context aware graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14333–14342.

[64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2921–2929.

[65] Zheming Zuo, Longzhi Yang, Yonghuai Liu, Fei Chao, Ran Song, and Yanpeng Qu. 2019. Histogram of fuzzy local spatio-temporal descriptors for video action recognition. *IEEE Transactions on Industrial Informatics* 16, 6 (2019), 4059–4067.