Applied Intelligence (2022) 52:2317–2331
https://doi.org/10.1007/s10489-021-02487-z

Predictively encoded graph convolutional network for noise-robust skeleton-based action recognition

Yongsang Yoon[1] · **Jongmin Yu**[1,2] · **Moongu Jeon**[1]

Accepted: 28 April 2021 / Published online: 9 June 2021

Abstract

In skeleton-based action recognition, graph convolutional networks (GCNs), which model human body skeletons using graphical components such as nodes and connections, have recently achieved remarkable performance. While the current state-of-the-art methods for skeleton-based action recognition usually assume that completely observed skeletons will be provided, it is problematic to realize this assumption in real-world scenarios since the captured skeletons may be incomplete or noisy. In this work, we propose a skeleton-based action recognition method that is robust to noise interference for the given skeleton features. The key insight of our approach is to train a model by maximizing the mutual information between normal and noisy skeletons using predictive coding in the latent space. We conducted comprehensive skeleton-based action recognition experiments with defective skeletons using the NTU-RGB+D and Kinetics-Skeleton datasets. The experimental results demonstrate that when the skeleton samples are noisy, our approach achieves outstanding performances compared with the existing state-of-the-art methods.

Keywords Predictive encoding · Graph convolutional network · Noise-robust · Skeleton-based action recognition

1 Introduction

Action recognition, which recognizes human behaviours using a computational system, is an important area in computer vision studies. Action recognition can be utilized in multiple applications, including industrial systems [48] and multimedia [53, 60]. Interest in this field has increased rapidly in recent years, and numerous studies have been proposed. Various modalities, such as appearance [11], depth [29], motion flow [51], and skeleton features [41] have been utilized to recognize human actions. Among the rapid advancements for learning useful representations automatically, various approaches have employed convolutional neural network (CNN) [8, 23] and recurrent neural network (RNN) [27, 59] models to learn spatiotemporal information and recognize human actions. These CNN and RNN based approaches, which take RGB images and motion flows (e.g., optical flow) as input, have achieved outstanding performances compared with earlier methods based on hand-crafted features (e.g. [55]). However, the drawback of deep learning approaches is that the learned representations may not be focused specifically on human actions because the entire areas of the video frames are provided when learning the representations [12, 39]. In contrast, skeleton features provide quantized information about human joints and bones. Compared to RGBs and motion flows, skeleton features can provide more compact and useful information in dynamic situations with complicated backgrounds [9, 12, 22].

Early approaches created skeleton data manually in the form of a sequence of joint-coordinate vectors [9, 27, 36, 42, 59] or as pseudo-images [22, 23, 30] and used the data to train RNNs or CNNs to predict the corresponding action classes. Intuitively, skeleton features can be represented as a graph structure because their components are homeomorphic. A graph based approach is

✉ Moongu Jeon
mgjeon@gist.ac.kr

Yongsang Yoon
nil@gist.ac.kr

Jongmin Yu
jm.andrew.yu@gmail.com

[1] School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, 61005, South Korea

[2] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, 610, WA, Australia

introduced in various methods such as action recognition [40] and 3D object retrieval [26] and has proven its effectiveness. In skeleton-based action recognition, the joints and bones of skeleton features can be defined as the vertices and connections of a graph, respectively. Recently, graph convolutional networks (GCNs) have achieved substantial success in skeleton-based action recognition [25, 39, 41]. ST-GCN [56] was the first work to use GCNs with a spatial approach to address skeleton models, and it showed impressive improvements. However, the spatial graph in ST-GCN is predefined and relies only on the physical structure of the human body. This makes it difficult to capture the relationships between closely related joints such as those between both hands in hand-related actions. To overcome this limitation, many methods [25, 37, 39, 41, 43] have been proposed that build adaptive graphs which pay dynamic attention to each joint based on the action being performed.

However, the existing approaches all assume that a complete skeleton is provided. Although recent studies on pose estimation [5, 50] and skeleton-feature construction [8, 22, 30] have shown precise performances, it is unrealistic to expect to obtain such perfect skeleton features without noise in real situations. Figure 1 shows the noisy skeleton estimated by AlphaPose [6]. Even though AlphaPose is a well-known pose estimation method, it results in noisy skeletons. To address this issue, Song et al. [43] defined a noisy skeleton as an '*incomplete skeleton*' in which some joints are spatially or temporally occluded. Song et al. [43] proposed a GCN-based method, named RA-GCN, that learns the distinctive features of currently unactivated (occluded) joints in multiple streams by utilizing class activation maps (CAM).

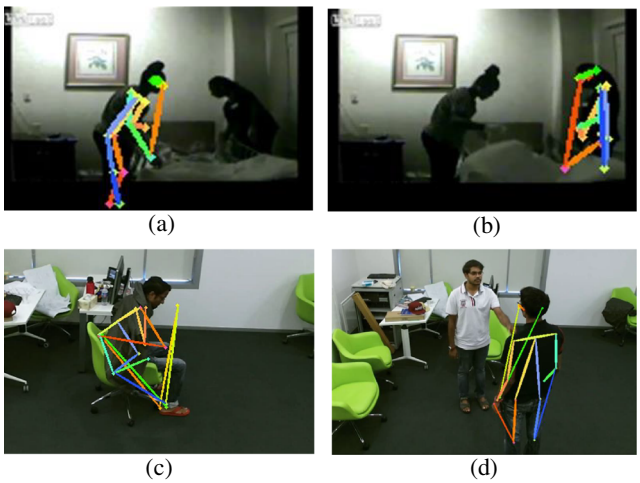

Fig. 1 Examples of noisy skeletons estimated by the existing pose estimation method (i.e., AlphaPose [6]). As shown in the figures, the estimated skeletons are corrupted because some of the estimated joints are in the wrong places. The snapshots were selected from the UCF-crime [45] and NTU datasets [36]

To the best of our knowledge, Song et al. [43] was the first GCN-based method to consider '*incomplete skeletons*'. Figure 2 shows some illustrations of occluded joints. Although regarding noisy joints as occluded joints is a reasonable approach, inaccurate joints should also be considered since inaccuracies can be easily observed when pose estimation is applied to real-world scenarios. In this work, therefore, we consider noisy skeletons rather than incomplete skeletons.

We present a predictively encoded graph convolutional network (PeGCN) model, which learns a noise-robust representation for skeleton-based action recognition. The key insight of our model is to learn such representations by predicting the perfect sample from a noisy sample in latent space via an autoregressive model that summarizes latent features and produces a feature context. We use a probabilistic contrastive loss to capture the most useful information and predict the perfect sample. To demonstrate the efficiency of the PeGCN on skeleton-based action recognition with noisy samples, we conducted various experiments using two datasets: NTU-RGB+D [36] and Kinetics-Skeleton [56]. The experimental results show that the PeGCN provides noise-robust action recognition performances using skeleton features and that its performance surpasses that of existing methods. The key contributions of our work are summarized as follows.

- We propose a general skeleton-based action recognition framework that is suitable for noisy skeletons generated from pose estimation. Any type of graph convolution network can be applied. To the best of our knowledge, only few methods have considered noisy skeletons.
- A simple yet effective network proposed with our framework, referred to as *Predicatively encoded graph convolutional network* (PeGCN), can derive complete skeleton feature from noisy skeleton feature in latent space by introducing predictive coding loss.
- We conducted extensive experiments with various setting and ablation studies. Our proposed approach outperforms existing methods in noise environments and is competitive in normal environments on public benchmark dataset Kinetics-skeleton [56] and NTU-RGBD [36].

The code has been made publicly available at https://github.com/andreYoo/PeGCNs.git. The remainder of this paper is organized as follows: In Section 2, we briefly review the related methods. In Section 3, we provide our motivation and intuition, as well as the structural details of our method, followed by the experimental results and analysis in Section 4 and conclusion in Section 5.

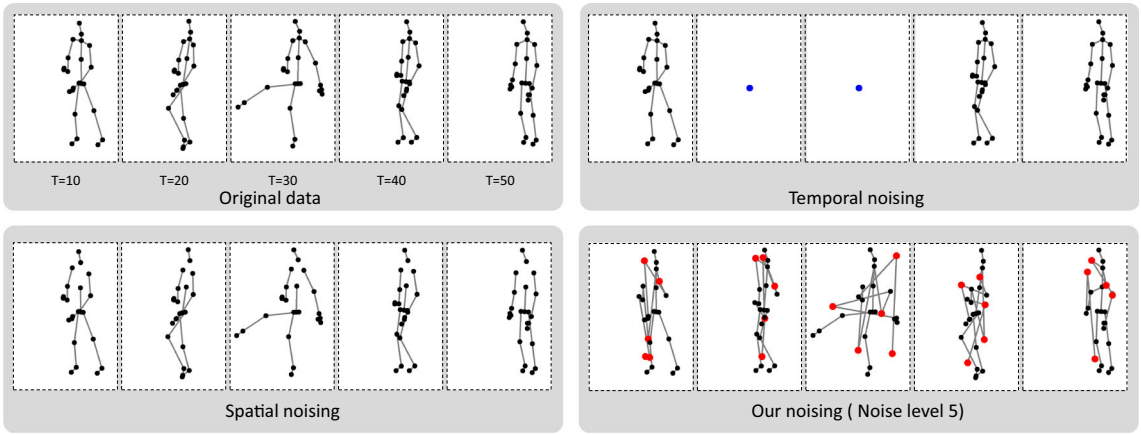

Fig. 2 Illustrations of various types of noisy skeletons, where T is the frame order associated with each skeleton: (*Original Data*) original skeleton samples;. (*Temporal noising*) and (*Spatial noising*) skeleton samples considered by Song et al. [43], which are spatially and temporally occluded; (*Our nosing*) a noisy skeleton sample generated by our noising approach using noise level

2 Related works

2.1 Skeleton-based Action recognition

The recent success of deep-learning techniques has had a significant impact on studies involving human action recognition. To model the spatiotemporal features of human actions, many works [27, 36, 40, 59] have attempted to extract appearance information with convolutional neural network (CNN) and temporal information with recurrent neural network (RNN) models. Recently, ST-GCN [56] successfully adopted a graph convolution network (GCN) to handle graphs in arbitrary forms; this was the first method to apply GCNs to skeleton-based action recognition.

The main drawback of GCN based methods is the spatial graph, which is predefined by only relying on the physical structure of the human body and is fixed to all GCN layers. To address these limitations, many methods [25, 37, 39, 41, 43] have been proposed for building adaptive graphs that pay attention dynamically to each joint based on the action being performed. The adaptive graph functions as a trainable mask that can learn the relationships between any joints, thus, increasing both flexibility and generality when constructing the graph. Shi et al. [39] proposed the 2s-AGCN model, which includes two adaptive graphs: a global graph and a local graph. Si et al. [41], in turn, combined an LSTM with a GCN (AGC-LSTM) to learn spatiotemporal representations from sequential skeletons, but most GCN-based models acquire temporal information with 1D convolution on the temporal axis. Spatial-based GCNs usually distribute graphs into multiple subgraphs using either distance partitioning or spatial configuration partitioning proposed in [56]. In contrast to these common partitioning strategies, Thakkar et al. [47] proposed a part-based GCN (PB-GCN) that learns the relationships between five body parts.

2.2 Noise-robust approach

Some works tried to handle the noise in the data which resulted in seriously penalizing the performance. Various types of data can be categorized as noise, such as occlusions, inaccurate positions, or outliers. In the image-based approach, one of the challenging issue is the occlusion in which two objects come too close together and seemingly merge or combine with each other. To handle this issue, Wang et al. [52] tried to learn the visibility of overlapped objects by inferring occlusion maps from a global detector using a combination of Histogram of Oriented Gradients (HOG) and Local Binary Pattern (LBP). Similarly, Gao et al. [13] introduced a binary pattern to propose occlusion-robust object detection by modeling segmentation-aware representation, which indicates whether pixels belong to the corresponding object or not. With this binary variable, a more compact and rich representation of the target object can be obtained by considering the target only. Wang et al. [54] argued that samples of occlusions and deformations are very rare, making it difficult to train, and proposed an alternative way which generates those samples by learning adversarial networks. In the skeleton-based approach, misaligned joints are regard as noise. Liu et al. [28] applied a global context attention module to the original LSTM in an attempt to concentrate on the salient joints while ignoring the irrelevant joints, rather than handling noise in each sample one by one. Song et al. [43] defined an '*incomplete skeleton*' which

contains noisy joints and tried to handle it by utilizing class activation maps (CAM).

2.3 Latent representation learning

In the area of vision, Latent representation learning is utilized in many ways including reducing computational cost or solving cross-domain problems. Gao et al. [15] utilized latent features to handle limited training samples in cross-domain cases by modeling pairwise network architecture with a self-attention mechanism. Also, a new public benchmark dataset for cross-domain action recognition (CDAR) was constructed. Moreover, Gao et al. [14] attempted to solve cross-domain few-shot action recognition by extracting efficient latent spatio-temporal dynamics from an attentive adversarial network. Carl et al. [49] tried to predict actions in video by anticipating the visual representation of the future frame from the current one. Rohit et al. [16] proposed a new representation which aggregates signals across time and space in the latent space for video-level encoding using the two-stream architecture. Rui et al. [35] proposed the Contrastive Video Representation Learning (CVPL) method to extract spatio-temproal representations from unlabeled videos. The representations are learned by using contrastive loss where positive samples are encoded to closer representations and negative samples are encoded to farther representations in latent space.

3 Predictively encoded graph convolutional networks

3.1 Motivation and intuition

Most of the existing skeleton-based action recognition methods only focus on complete skeleton data which are captured in the constrained environment with a depth camera. Even if the depth camera has an advantage over capturing 3D data, noise can still be contained in the data due to geometric conditions such as camera viewpoint, or active objects. The situation will be further aggravated if joints are estimated from the usual RGB camera. Such noise will cause severe performance degradation even for the *state-of-the-art* action recognition methods, especially in real-world scenarios which likely contain even more noise.

To solve this problem, we first define a *noisy-skeleton* by assigning random noise joints in the complete skeleton. The noisy joints addressed in this paper are misaligned joints compared to their actual positions. They are generated by relocating the joints to random positions within the bounding box of a person. Song et al. [43] also defined an '*incomplete skeleton*', in which the joints are spatially or temporally occluded skeleton samples, by replacing the joints to a single fixed position. We assume that generating noise from the bounding box is more realistic since joints are estimated within the bounding box in a top-down manner pose-estimation. Such noise patterns are inherently unpredictable. Therefore, modeling noise information explicitly in data-driven approaches is impractical.

To develop a precise action recognition method, it is important to not only learn a global representation but also have outstanding generalizability in order to be robust to diverse types of noise. Deep learning is well known as an effective way to improve model generalizability for various visual recognition studies [2, 9, 17]. GCN is a unified framework consisting of a graph structure and deep learning; therefore, it also has the advantage of improved generalization performance. Based on this advantage, the dominant approach to training skeleton-based action recognition methods based on GCNs is to initially extract information from skeleton samples using GCNs and then to compute the unimodal loss e.g. cross entropy [25, 37, 39, 41, 43, 56]. This approach can be regarded as a direct end-to-end learning approach such as modeling $p(\bar{o}|x)$ between skeleton samples x and a corresponding acting label $\bar{o}$. However, this approach is computationally intensive and a waste of the representation capacity of the model. For example, slight noise, which could be alleviated during generalization via a nonlinear network structure, does not need to be considered as meaningful. Therefore, to derive an optimal global representation, it may not be appropriate to derive a mapping model $p(\bar{o}|x)$ directly.

The key insight underlying the developed PeGCN for noise-robust skeleton-based action recognition is to learn encoded representations of the underlying shared information. These representations can be obtained by predicting the missing information between a complete sample and a noisy sample in the latent space. Using this approach, we can recognize actions using an encoded complete skeleton instead of a noisy skeleton, since incompleteness has severe negative impacts on the recognition performance. Utilization of the mined core joints was also considered since each joint contributes differently depending on the action performed. However, according to ST-GCN [56], the movement of the upper-body may not be negligible even when a leg-related action (e.g., kicking) is performed. Using only a small set of joints may not be enough to fully understand the action. Thus we try to exploit the shared information rather than mining the core joints. This idea is inspired by predictive coding [1, 10, 33], which is one of the oldest techniques in signal processing for data compression. Predictive coding has recently been applied to unsupervised learning in order to learn word representations [31] by predicting neighbouring words. The latent space approach has

the following advantages. First, an action recognition model needs to infer more global structures since it requires relatively longer time samples compared to other tasks, such as event detection [57, 58] or change detection [19]. When inferring the global structure, high-level information (i.e., latent space) is more suitable than low-level information. Second, the recognition performance is more likely to be seriously affected by the global noise in the latent representation than the local noise, which can be reduced via deep learning nonlinear weighted kernel structures.

To predict appropriate information from noisy skeleton features, we train the model to maximally conserve the mutual information (MI) between the two inputs. The mutual information, which can measure the mutual dependence between the two inputs, is defined as follows:

$$I(x;\bar{\alpha}) = \sum_{x,\bar{\alpha}} p(x,\bar{\alpha}) \log \frac{p(x|\bar{\alpha})}{p(x)} \tag{1}$$

where x and $\bar{\alpha}$ are complete skeleton and noisy representation, respectively. By maximizing the mutual information between the two representations (which are bounded by the MI between the inputs), we extract the underlying latent variable, which is robust to the global noise.

3.2 Structural details

Figure 3 illustrates the PeGCN training and inference pipeline. The PeGCN consists of GCN module f_{enc} and autoregressive module f_{ar}. In preprocessing, the noisy skeleton sample x' is generated from the given skeleton sample x. The GCN module f_{enc} encodes skeleton samples $x*$ into a latent space $\alpha*$, where $*$ indicates the input type: a normal type (x and α) or noisy type (x' and α'). The autoregressive module f_{ar} summarizes the latent representation and produces the contextual latent representation $\bar{\alpha} = f_{\text{ar}}(\alpha')$. Note that the autoregressive model is specialized in handling time series data better than a common linear transformation.

As discussed in the previous section, we do not predict the appropriate information directly from the noisy skeleton via the generative model $p(x|\bar{\alpha})$. Instead, we utilize a density ratio [33], which helps preserve the mutual information between two representations, as follows:

$$D(x;\bar{\alpha}) = \frac{p(x|\bar{\alpha})}{p(x)} \tag{2}$$

where x and $\bar{\alpha}$ denote the skeleton sample and the contextual latent representations, respectively. By combining the encoded representation $\bar{\alpha}$ and the density ratio $D(x;\bar{\alpha})$, model is alleviated from modeling high dimensional distribution x. Even though we cannot derive $p(x)$ or $p(x|x')$ directly, we can use the samples from these distributions, which also allows the application of well-known techniques (e.g. important sampling [3] and noise contrastive estimation [18, 20, 32]).

The backbone network for our GCN module f_{enc} is the GCN part of Js-AGCN [39]. The GCN module is composed of adaptive graph convolutional layers, which optimize the graph topology in combination with the other parameters of the network in an end-to-end learning manner. The adaptive convolutional layer is defined by

$$\boldsymbol{f}_{\text{out}} = \sum_{k}^{K_v} \boldsymbol{W}_k \boldsymbol{f}_{\text{in}} (\boldsymbol{A}_k + \boldsymbol{B}_k + \boldsymbol{C}_k), \tag{3}$$

where $\boldsymbol{A}_k$ is the original normalized adjacency matrix, $\boldsymbol{B}_k$ is a global attention matrix and $\boldsymbol{C}_k$ is an individual attention matrix which is a unique graph for each sample. With (3), the latent representation of a given skeleton can be obtained. We employed the GCN from the 2s-AGCN model, with the exception of fully connected networks.

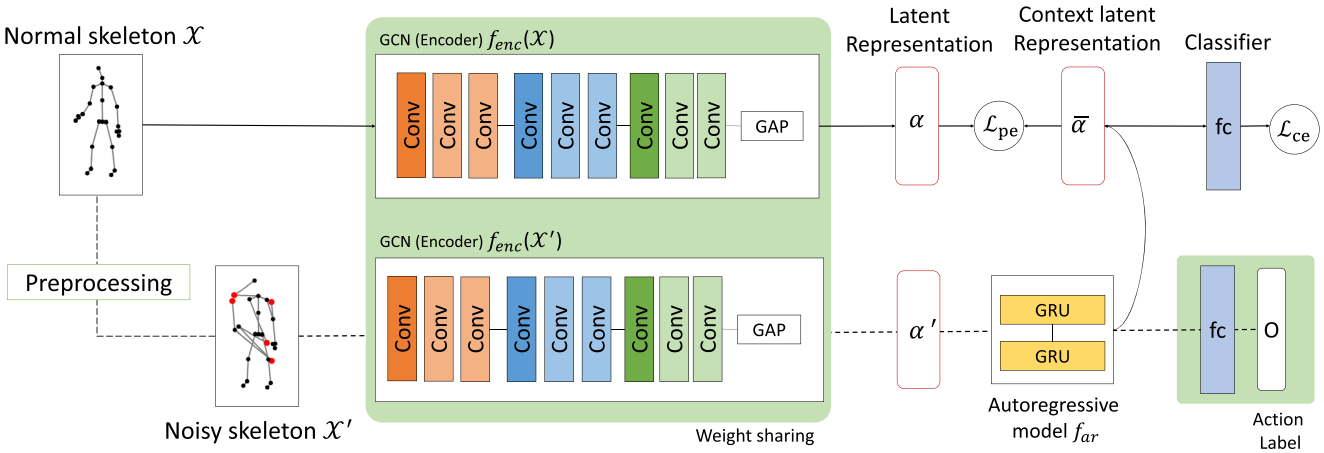

Fig. 3 PeGCN has two branches for normal and noisy represented in solid-line and dotted-line, respectively. Both branches are used in training while only the noisy branch is used in testing. *GAP* and *fc* denote global average pooling and fully connected layer, respectively. Note that the noisy skeleton is generated by *preprocessing* at every run time

RNNs with gated recurrent units (GRUs) [7] were used for the autoregressive module f_{ar}. This selection can easily be replaced by other linear transformations or nonlinear networks. Note that any type of GCN model or autoregressive model can be applied in the proposed method. It seems likely that more recent advancements in GCNs and autoregressive models could achieve better results.

3.3 Training and inference

Both the GCN and autoregressive modules are jointly trained to optimize the loss and maximize the MI between two latent representations of normal and noisy skeleton features, which we call predictive encoding loss. With a given set $X \in x_1, ..., x_N$ which contains one positive sample $p(x_*|\bar{\alpha})$ and $N-1$ negative samples from the distribution $p(x')$, the following loss is optimized:

$$\mathcal{L}_{\text{pe}} = -\mathbb{E}_X \left[\log \frac{D(x_*; \bar{\alpha})}{\sum_{x \in X} D(x; \bar{\alpha})} \right], \tag{4}$$

where x_* denotes the positive sample corresponding to $\bar{\alpha}$. Note that (4) is the categorical cross entropy. According to Oord et al. [33], optimizing this loss will result in estimating the density ratio in (2). In other words, minimizing the loss $\mathcal{L}_{\text{pe}}$ will lead to maximizing the mutual information $I(x; \bar{\alpha})$. Action recognition should identify the action class of a given skeleton sample. Using $\mathcal{L}_{\text{pe}}$ alone cannot achieve this goal since it is only focused on maximizing the MI between two latent representations. Therefore, similar to other studies [38, 39, 43], the cross entropy loss is exploited as follows:

$$\mathcal{L}_{ce} = -\sum_{i}^{C} \bar{o}_i \log(o_i), \tag{5}$$

where C is the number of action classes, $\bar{o}$ is the given annotation for an action sample, and o is the output of the fully connected network for the classification task in the inference stage. Consequently, to train the noise-robust skeleton-based action recognition model, the total loss functions are straightforwardly defined by the sum of the cross entropy loss $\mathcal{L}_{\text{ce}}$ and the proposed predictive loss function with the balancing weight λ, which is represented as follows:

$$\mathcal{L}_{\text{ae}} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{pe}}. \tag{6}$$

In all our experiments, the best performance achieved when λ is set to 0.1. Action recognition using the PeGCN is straightforward. In the test step, the GCN module f_{enc} encodes an input skeleton sample into the latent space, and the autoregressive model f_{ar} summarizes the latent feature and generates the context latent representation $\bar{\alpha}$. Finally, $\bar{\alpha}$ is used as the input to the fully connected network for action recognition (Fig. 3).

4 Experiments

4.1 Experimental setting

To evaluate the action recognition performance, the PeGCN is tested on two datasets: NTU-RGB+D dataset [36] and Kinetics-Skeleton [56]. We followed the same evaluation protocol described in [39] in which the Top-1 and Top-5 recognition accuracies are evaluated. The details of each dataset are as follows:

The NTU-RGB+D dataset [36] is one of the largest datasets in skeleton-based action recognition, containing approximately 56,000 samples and consisting of 60 indoor activities (e.g., hand clapping or drinking water). The samples were captured by Microsoft Kinect v2 at three different angles (-45, 0, 45) with 40 volunteers. In the skeleton sequences, 3D spatial coordinates (X, Y, Z) for 25 joints are provided for each human action. This dataset has two benchmark protocols: cross-view (CV) and cross-subject (CS). In the CV protocol, the samples are split into training and test sets according to the camera angle with 37,920 and 18,960 samples, respectively. In the CS protocol, the samples are split into training and test sets based on the subjects that appear in the sequences. Some subjects are designated as training samples, and the remaining subjects are designated as test samples. Under the cross-subject protocol, the training and test subsets contain 40,320 and 16,560 samples, respectively. Following these protocols, the top-1 accuracy scores on both benchmarks are reported.

The Kinetics-Skeleton dataset [56] is another large-scale skeleton action dataset generated from the Kinetics dataset [21], which contains 34,000 video clips collected from YouTube that have a wide variety of characteristics such as illumination changes and background color variations. Each video clip is labeled from a set of 400 action classes. The skeleton model is estimated with the publicly available OpenPose toolbox [4], which yields 2D locations and 1D confidence scores for 18 joints. The top two people with the highest average joint confidence scores in the video clips are selected when multiple people exist in the scene. The length of each skeleton sequence is fixed at 300 by repeating

or sampling the sequence. The dataset (Kinetics-Skeleton) contains 240,000k samples for the training set and 20,000k samples for the validation set.

A noisy skeleton can be defined as a skeleton that contains inaccurate joint positions. A joint position is determined within the person area in top-down pose estimation. Based on this background, a noisy skeleton is generated as follows: 1) The noise level is determined manually. 2) Based on the noise level, some joints are selected randomly. 3) A boundary box is determined by all the joints in the complete skeleton, as described in Figure 4a. 4) A random point in the boundary box is assigned to each selected joint in every frame. Figure 4b describes how random points are assigned to the existing joints depending on the noise level. In this manner, we can generate noisy samples based on the assumption that the skeleton is estimated in a top-down manner where one finds a person first and then estimates joints from the person area. Additionally, the generated noisy skeleton is similar to those estimated by OpenPose from real-world videos. Note that temporal noise is not considered in this work. Only spatial noise is addressed since we believe that spatially disordered joints are the more common noise.

The common hyperparameter settings used to train the PeGCN are as follows. The number of epochs is set at 50 and 65 for the NTU-RGB+D and Kinetics-Skeleton datasets, respectively. Given that our computational resources are limited, the batch size was reduced to 32, which is half of the original batch size of our backbone network [39] and which can negatively affect the PeGCN's action recognition performance. The stochastic gradient descent and weight decay are utilized as optimization algorithms. The experiments are divided into two parts: the first involves the ablation study, and the second performs comparisons with existing state-of-the-art methods.

4.2 Ablation study

4.2.1 Experimental protocol

The performance analysis was conducted based on the hyperparameter settings of the PeGCN. The hyperparameters that most significantly affect the action recognition performance of the PeGCN are the noise level and the composition of the loss function. The performance analysis based on the noise level and the loss function composition settings in the training step is carried out as follows. First, two PeGCN models trained by $\mathcal{L}_{\text{ce}}$ (PeGCN$_{\text{_ce}}$) and $\mathcal{L}_{\text{ae}}$ (PeGCN$_{\text{_ae}}$) are constructed, and then, each model is trained with 1, 3, and 5 noise levels. The other parameters are set exactly the same as the parameter settings mentioned in the preceding section. Next, these models are evaluated with noise levels between 1 and 13. Finally, the trends of the cross entropy losses and the predictive coding losses of these models are observed, and their action recognition accuracies are compared. For efficiency, the ablation study used only the CV protocol of the NTU-RGB+D dataset.

4.2.2 Experimental results

Table 1 shows that the action recognition accuracy depends on the noise level and the loss function settings in the training step. The best accuracy was achieved by $PeGCN_{\text{\_ae\_}n5}$ in all test cases. Specifically, an accuracy of 93.33 and 90.28 was achieved for noise level N3 and N13, respectively. The models trained at different noise levels, $PeGCN_{\text{\_ce\_}n1}$, $PeGCN_{\text{\_ce\_}n3}$ and $PeGCN_{\text{\_ce\_}n5}$, achieved an accuracy of 85.88, 89.61 and 89.79 at testing level N7, respectively. The performance improved as the training noise level increased, when the total number of joints is 25. At testing level N10, $PeGCN_{\text{\_ae\_}n5}$ trained with

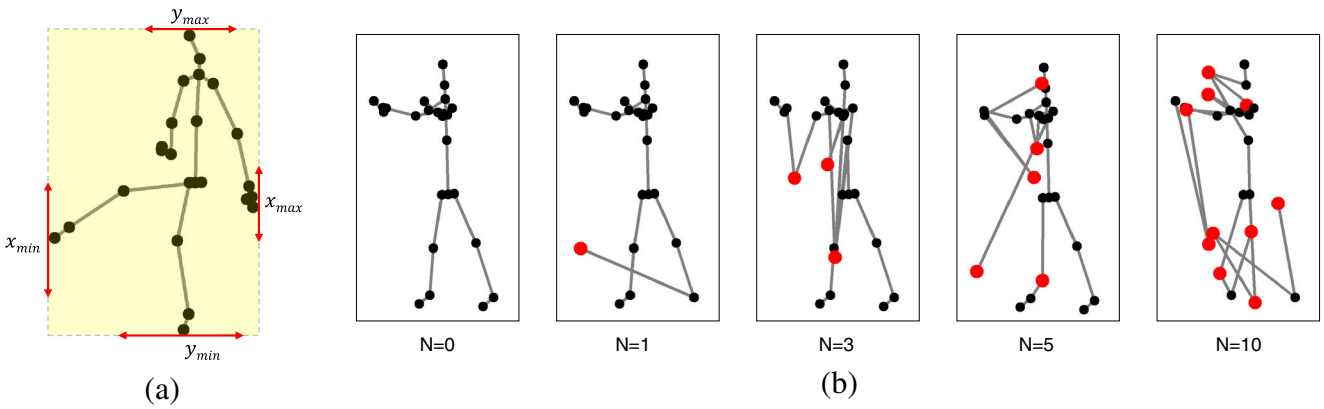

Fig. 4 Illustrations of setting the candidate scope to generate noisy joints and example samples of noisy skeleton depending on the noise level: **a** defining the scope to generate noisy joints using a given skeleton sample; **b** noisy skeleton samples created from an original sample depending on the noise level

Table 1 The top-1 accuracy of the PeGCN depends on the loss function and the noise level setting under the Cross-View (CV) protocol of the NTU-RGB+D dataset

Models	Test level				
	N3	N5	N7	N10	N13
w/o the predictive encoding loss $\mathcal{L}_{\text{pe}}$					
$PeGCN_{\_ce\_n1}$	91.52(±0.21)	89.33(±0.25)	85.88(±0.15)	78.47(±0.19)	66.66(±0.15)
$PeGCN_{\_ce\_n3}$	92.64(±0.26)	91.1(±0.21)	89.61(±0.25)	85.16(±0.25)	77.99(±0.19)
$PeGCN_{\_ce\_n5}$	92.15(±0.14)	91.33(±0.29)	89.79(±0.21)	87.06(±0.19)	82.23(±0.21)
w/ the predictive encoding loss $\mathcal{L}_{\text{pe}}$					
$PeGCN_{\_ae\_n1}$	91.89(±0.27)	90.79(±0.12)	89.61(±0.31)	86.9(±0.21)	83.28(±0.16)
$PeGCN_{\_ae\_n3}$	92.71(±0.19)	92.34(±0.31)	91.62(±0.21)	90.41(±0.14)	88.96(±0.14)
$PeGCN_{\_ae\_n5}$	**93.33**(±0.12)	**92.22**(±0.18)	**92.13**(±0.22)	**91.24**(±0.24)	**90.28**(±0.31)

The model name is determined by the loss function and the noise level. For example, the model $PeGCN_{\_pe\_n5}$ is the model which is trained with loss $\mathcal{L}_{\text{pe}}$ at noise level 5. The results are denoted in the form of mean accuracy and standard deviation in the parentheses. The boldface figures indicate the highest performance for each experiment, respectively

$\mathcal{L}_{\text{ae}}$ achieved a higher accuracy of 91.24, compared to $PeGCN_{\_ce\_n5}$ which achieved 87.06, when both models are trained at the same noise level. These overall quantitative results demonstrate that among the models trained at the same noise level, the model trained with the total loss function $\mathcal{L}_{\text{ae}}$ usually performs better. It also suggests that performance degradation occurs much faster in PeGCNs trained using only the cross entropy loss, compared with other models.

Moreover, the training losses of each model show the efficiency of $\mathcal{L}_{\text{pe}}$ when learning the noise-robust representation. Figure 5a illustrates the changes in the cross entropy losses for PeGCN$_{_ce}$ and PeGCN$_{_ae}$ trained at different noise levels. Although both the CE losses of PeGCN$_{_ce}$ and PeGCN$_{_ae}$ are similar by the end of training, PeGCN$_{_ce}$ initially converges much faster than does PeGCN$_{total}$ in all cases. This trend suggests that training with only $\mathcal{L}_{\text{ce}}$ makes it easier to convert to a poor locally optimized solution than training $\mathcal{L}_{\text{ce}}$ with $\mathcal{L}_{\text{pe}}$, based on their respective accuracies. Figure 5b illustrates the predictive coding loss of PeGCN$_{_ae}$ at different noise levels.

Interestingly, the coding loss of the PeGCN$_{_ae}$ trained at noise level 1 is relatively higher than that of the PeGCN$_{_ae}$ trained at noise level 3 or 5. These trends can be attributed to the difficulties of making generalizations from a small volume of noise, where the accuracy of PeGCN$_{_ae}$ trained at noise level 1 is lower than that of others (see Table 1). In the training step, a higher noise level can provide more diversity in the training samples than can a lower noise level. Consequently, the ablation study demonstrates that adopting a higher noise level in the training step can improve the action recognition performance in the test step,

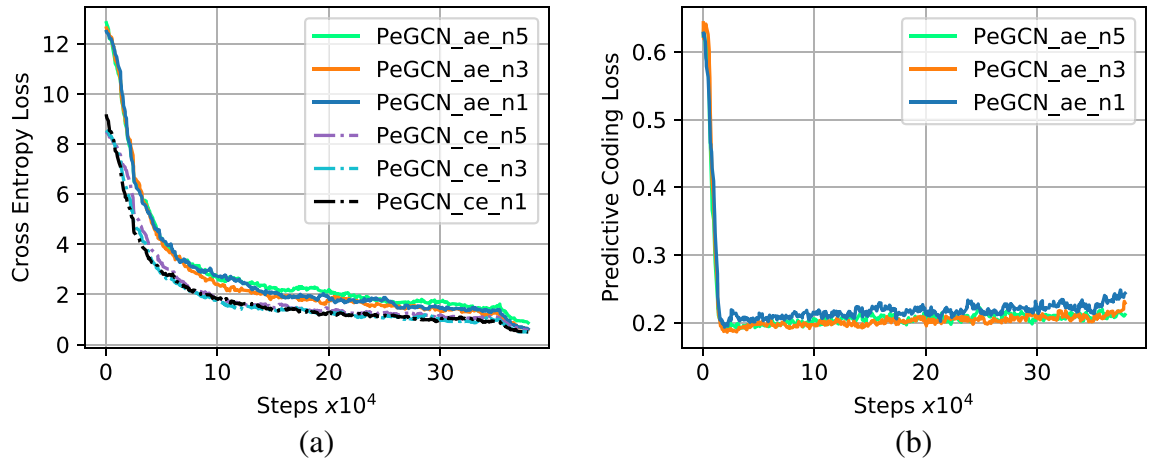

Fig. 5 Trends of the cross entropy and predictive encoding losses according to the PeGCNs trained at different conditions under the CV protocol of the NTU-RGB+D dataset: **a** represents the curves of cross entropy functions $\mathcal{L}_{\text{ce}}$; and **b** represents the curves of the predictive encoding losses $\mathcal{L}_{\text{pe}}$. Note that both graphs are smoothed for better readability

but not in a linearly proportional fashion. For experimental efficiency, further studies comparing the PeGCN with the existing state-of-the-art methods are conducted only with $\text{PeGCN}_{\text{total}}$ trained at noise level 5.

4.2.3 Experimental protocol

We evaluated the PeGCN on both normal and noisy skeleton samples by following the general experimental protocol described in the NTU-RGB+D dataset [36] and the Kinetics-Skeleton dataset [21]. For both datasets, the top-1 and top-5 accuracies were computed for the performance comparison. In the experiments on the NTU-RBGD dataset, both the cross-view (CV) and cross-subject (CS) protocols were applied. To reduce the volatility of performance due to the randomness of noised joints, all the experiments were iteratively conducted 10 times, and the mean accuracies were used for the comparison.

We compared the PeGCN with several recently proposed state-of-the-art methods. For experimental efficiency and to ensure fair comparisons, the methods that were proposed before 2018, or whose performance is at least 5% lower than ours on normal skeleton evaluation (e.g. [9, 12, 27, 36, 59]) were excluded from this comparison (see Table 2). In particular, in the experiments using noisy skeleton samples, the methods whose source codes were not released by the paper authors were also excluded from the experiments [34, 37]. In addition, some methods whose source codes have been made public were also excluded based on the following criteria: 1) the source code was released but not by the original authors; 2) the paper has yet to be officially published in a journal or a conference proceeding.

Table 2 Performance comparisons on complete skeletons from the NTU-RGB+D and Kinetics-Skeleton datasets

Methods	Architecture	NTU-CS		NTU-CV		Kinetics-Skeleton	
		Top1	Top5	Top1	Top5	Top1	Top5
Feature Enc [12]	Hand-crafted	–	–	–	–	14.9	25.8
HBRNN [9]	RNN	59.1	–	64.0	–	–	–
Deep LSTM [36]	LSTM	60.7	–	67.3	–	16.4	35.3
ST-LSTM [27]	LSTM	69.2	–	77.7	–	–	–
STA-LSTM [42]	LSTM	73.4	–	81.2	–	–	–
VA-LSTM [59]	LSTM	80.7	–	88.8	–	–	–
TCN [23]	CNN	74.3	–	83.1	–	20.3	40.0
Clips+CNN+MTLN [22]	CNN	79.6	–	84.8	–	–	–
Synthesized CNN [30]	CNN	80.0	–	87.2	–	–	–
3scale ResNet152 [24]	CNN	85.0	–	92.3	–	–	–
DPRL+GCNN [46]	GCN	83.6	–	89.8	–	–	–
AGC-LSTM(Joint&Part) [41]	GCN+LSTM	89.2	–	95.0	–	–	–
AS-GCN [25]	GCN	86.8	–	94.2	–	34.8	56.5
ST-GCN* [56]	GCN	81.6 (81.5)	96.9	88.8 (88.3)	98.8	31.6 (30.7)	53.7 (52.8)
2s RA-GCN* [43]	GCN	85.8 (85.8)	98.2	93.0 (93.0)	99.3	–	–
3s RA-GCN* [43]	GCN	85.9 (85.9)	98.1	93.5 (93.5)	99.3	–	–
PB-GCN* [47]	GCN	87.0 (87.5)	98.3	93.4 (93.2)	99.4	–	–
Js-AGCN* (Backbone) [39]	GCN	85.4	97.3	93.1 (93.7)	99.08	34.4 (35.1)	57.1 (57.1)
Bs-AGCN* [39]	GCN	87.0	97.5	94.1 (93.2)	99.23	34.1 (33.3)	57.0 (55.7)
2s-AGCN* [39]	GCN	88.8 (88.5)	98.1	95.3 (95.1)	99.4	36.8 (36.1)	59.2 (58.7)
GCN-NAS(Joint&Bone) [34]	GCN	89.4	–	95.7	–	37.1	60.1
DGNN [37]	GCN	89.9	–	96.1	–	36.9	59.6
JB-AAGCN [38]	GCN	89.4	–	96.0	–	37.4	60.4
MS-AAGCN [38]	GCN	**90.0**	–	**96.2**	–	**37.8**	**61.0**
PeGCN_{ae}	GCN	85.6	96.79	93.9	99.02	34.0	57.24

NTU-CV and NTU-CS are evaluation protocol Cross-View (CV) and Cross-Subject (CS) of NTU-RGB+d dataset. Each protocol is described in Section Experimental setting Section 4.1 A hyphen ('-') indicates that the results were not reported. The symbol * indicates a model trained by the authors of this paper, and figures in the parentheses represent the reported accuracy. The boldface numbers denote the highest performance for each experiment

4.2.4 Experiment with normal skeletons

Initially, we compared the PeGCN with other existing state-of-the-art methods on normal skeleton samples. For experimental consistency, several methods were tested using the publicly available source codes compiled by the authors of this paper [39, 43, 47, 56]. Table 2 contains the top 1 accuracies for the CS and CV protocols on the NTU-RGB+D dataset and the top 1 and top 5 accuracies on the Kinetics dataset. In the experiments, the PeGCN achieves an accuracy of 85.6 and 93.9 for the CS and CV protocols of the NTU-RGB+D dataset, respectively. The PeGCN produces 34.0 and 57.2 for the top 1 and top 5 accuracies on the Kineitcs-skeleton dataset. The MS-AAGCN [38] achieves state-of-the-art performances—90.0 for the CS protocol and 96.2 for the CV protocol. The second-highest performance is achieved by the DGNN [37], resulting in 89.9 and 96.1 for the CS and CV protocols, respectively. On the Kinetics-Skeleton dataset, the MS-AAGCN [38] scores 37.8 for the top-1 and 61.0 for the top-5. The MS-AAGCN also scores the second-highest performance on this dataset with 37.8 and 61.0 for the top-1 and top-5, respectively. In comparison, the PeGCN produces performances that are better than or comparable to several of the other methods. The Js-AGCN [39], which is used as the backbone network for PeGCN $_{ae}$, achieves the respective accuracy of 85.4 and 93.1 for the CS and CV protocols on the NTU-RGB+D dataset, which is only slightly below ours.

Nevertheless, the performance of PeGCN$_{ae}$ is generally lower than that of a few other methods, such as MS-AAGCN [38], DGNN [37], GCN-NAS [34], and AS-GCN [25]. The performance gap between these state-of-the-art methods and the PeGCN can be interpreted as follows: The MS-AAGCN [38] uses additional attention modules (e.g. spatial, temporal, channel-wise attention) and exploits four different modalities, including joint and bone information and motion information. During training, its batch size is twice that of ours, and the adaptive graphs are fixed in the first 5 epochs to achieve better learning, as explained in the DGNN [37]. The MS-AAGCN achieved top-1 accuracies higher than ours by 4.4%, 2.3% and 3.8% on CS, CV and Kinetics, respectively. Although the DGNN [37] uses the same batch size (32), it has a longer training epoch (120), while our training epoch is 50 for NTU-RGB+D and 65 for Kinetics Skeleton. In addition, the DGNN utilizes both joint and bone information through a directed acylic graph. This leads to improvements in its top-1 accuracy of 4.3%, 2.2% and 2.9% over the PeGCN on CS, CV and Kinetics, respectively. Other methods (such as GCN-NAS [34] and AS-GCN [25]) also adopt longer training epochs than ours, and their learning rate decays more frequently.

4.3 Comparisons with the existing state-of-the-art methods

4.3.1 Experiment with noisy skeletons

The experimental results on skeleton-based action recognition with noisy samples clearly demonstrate the efficiency of PeGCN when recognizing actions on noisy skeleton samples. In contrast to the other approaches, in which the performance rapidly degrades when the noise levels increase, the PeGCN shows noise-robust action recognition performances. As shown in Table 3a of the results on the CV protocol using the NTU-RGB+D dataset, the PeGCN achieves accuracies of 93.75 and 91.84 which are the highest accuracies obtained in experiments with noise levels 1 and 10. No other method has reached 88% accuracy level even at noise level 1. Js-AAGCN* [38] produces an accuracy of 87.51 for noise level 1. However, its recognition performance degrades steeply as the noise level increases. In the experiment with noise level 10, the performance of Js-AAGCN* is only 51.38. Furthermore as shown in the Fig. 6, all methods except our PeGCN tends to drop accuracy significantly as noise level increased. PeGCN showed considerable performance that maintains 90% of accuracy in highly noised environment (i.e., noise level 13).

The experimental results for the CS protocol using the NTU-RGB+D dataset also suggest that the PeGCN can provide more noise-robust performance than the existing state-of-the-art methods. As shown in Table 3b, the PeGCN achieves accuracies of 85.18 and 84.54 in the experiments where the noise levels are 1 and 5, respectively. The performance gap between these two figures is less than 1%, which is significantly lower than the performance gap in the other methods. 2s-AGCN et al. [39], who achieved a state-of-the-art performance in experiments with normal skeleton samples (see Table 2), reported accuracies of 76.37 and 49.41 in noise-1 and 5 experiments, respectively, and the gap between these two accuracies is greater than 26%. In the experiments with noise level 10, the performances of the other methods are all lower than 50, while PeGCN$_{total}$ obtains an accuracy of 83.13.

The experimental results on the Kinetics-Skeleton dataset also show noise-robust performances on noisy skeletons. As shown in Table 3c, the PeGCN obtains better accuracies than the other methods for all noise level cases. The Js-AGCN [39] achieved top-1 and top-5 accuracies of 23.04 and 43.45 at noise level 1, respectively. However, its performance drops dramatically when the noisy level increases, resulting in a top-1 accuracy of only 11.05 and 6.44 in noise-5 and noise-10 experiments, respectively. The performance dropped more than 15% compared to that in the noise-1 experiment. Although the accuracy of PeGCN

Table 3 Recognition accuracies depending on the noise level on the NTU-RGB+D and Kinetics dataset

Methods	Noise level							
	1		3		5		10	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
(a) NTU-RGB+D Cross-View (CV) Accuracy								
ST-GCN* [56]	83.09	97.27	76.26	94.75	68.89	90.76	51.77	78.08
PB-GCN* [47]	79.67	95.17	67.11	88.27	54.94	80.12	35.57	60.86
3s RA-GCN* [43]	79.85	92.63	66.55	84.76	55.74	76.51	35.83	58.28
2s RA-GCN* [43]	79.57	92.76	66.23	84.23	54.53	74.84	35.04	55.01
Js-AGCN* [39]	86.19	96.26	77.23	92.35	68.71	87.19	48.46	71.05
Js-AAGCN* [38]	87.51	95.96	79.84	91.77	71.51	85.9	51.38	68.91
Bs-AGCN* [39]	87.63	97.91	79.23	94.81	71.03	90.06	50.91	74.03
2s-AGCN* [39]	89.34	98.48	83.46	96.23	76.81	92.74	56.91	78.47
PeGCN (ours)	**93.75**	**99.15**	**93.33**	**99.11**	**92.92**	**99.04**	**91.84**	**98.92**
(b) NTU-RGB+D Cross-Subject (CS) Accuracy								
ST-GCN* [56]	73.1	93.26	64.82	88.23	56.8	82.57	40.5	68.17
PB-GCN* [47]	77.17	94.81	67.03	89.35	56.86	81.88	37.78	64.15
3s RA-GCN* [43]	71.52	91.23	57.06	80.86	46.82	70.52	28.87	50.51
2s RA-GCN* [43]	72.29	90.16	58.29	79.35	47.12	69.18	29.45	49.15
Js-AGCN* [39]	76.01	92.02	65.44	84.87	54.49	76.27	35.82	56.95
Js-AAGCN* [38]	80.57	93.5	73.17	89.37	66.13	84.19	49.11	70.15
Bs-AGCN* [39]	79.35	93.98	70.29	89.15	61.05	83.41	43.95	68.94
2s-AGCN* [39]	83.02	96.12	75.12	92.32	66.65	87.01	47.33	70.84
PeGCN (ours)	**85.18**	**97.08**	**85.01**	**96.85**	**84.54**	**96.95**	**83.13**	**96.39**
(c) Kinetics-Skeleton Accuracy								
ST-GCN* [56]	22.17	42.59	8.92	22.19	3.54	11.17	0.92	3.8
Js-AGCN* [39]	23.04	43.45	15.41	32.01	11.05	24.67	6.44	15.2
Js-AAGCN* [38]	27.38	48.5	19.26	37.37	14.03	28.8	7.89	17.42
Bs-AGCN* [39]	24.11	45.56	16.14	33.38	11.95	25.63	6.67	15.9
2s-AGCN* [39]	28.32	49.11	20.05	38.14	14.89	29.47	8.72	18.04
PeGCN (ours)	**33.23**	**55.6**	**32.78**	**55.22**	**32.39**	**54.34**	**29.63**	**51.78**

The symbol * indicates methods trained and tested by the authors of this paper. The numbers in boldface denote the highest performance for each experiment

also decreases as the noise level increases, the accuracy dropoff rate is much less than that of the other models. At noise level 1, PeGCN_{ae} achieved 33.23 accuracy, while the other methods achieved less than 29%. Moreover, at noise level 5, the other methods showed an accuracy below 15%, while the PeGCN_{ae} achieved 32.39.

To show the effectiveness of our framework, a comparison is conducted between accuracies obtained with and without the application of predictive encoding learning. As shown in the Fig. 7, all three methods are dramatically improved on both protocols (i.e., CV and CS) and the performance decline rates are considerably alleviated. It is interesting to observe that ResGCN [44], to which residual connection is applied, outperformed ST-GCN [56] on the normal skeleton data but showed relatively low performances under the noisy environment.

4.4 Analysis and discussion

The overall results indicate that the PeGCN provides outstanding skeleton-based action recognition that is robust to noisy samples, compared to the existing state-of-the-art methods. The accuracy scores achieved by the PeGCN for all noise levels on the NTU-RGB+D and Kinetics datasets outperform those of the compared models. In addition, the performance gap between the PeGCN and other methods is proportional to the noise level. In the experiment on noise level 10, the performances of nearly all

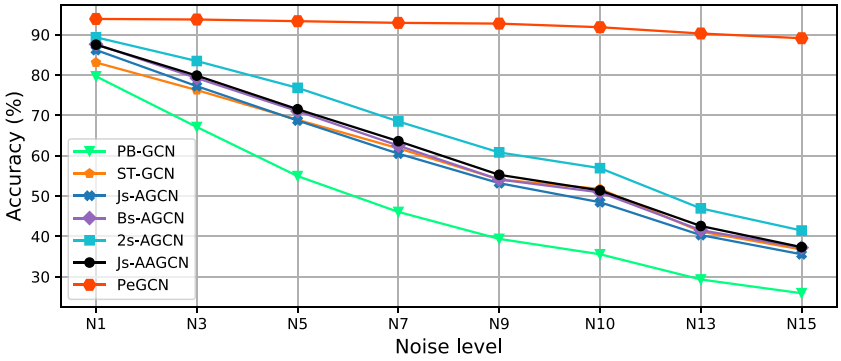

Fig. 6 Illustration of performance dropoff of each method. The results was tested on noise level between 1 to 15 under NTU-RGB+D cross-view (CV) protocol

the methods except PeGCN degrade by over 30% compared with their results on normal samples. In addition to the accuracy scores, the standard deviations also suggest that the PeGCN has advantages in noise-robust skeleton-based action recognition.

Interestingly, among the experimental results, the RA-GCN [43], which was proposed to recognize actions from incomplete skeletons, achieves relatively worse accuracy scores (Table 3a and b) than do other methods [37, 39, 56] that do not consider skeletons with noise information. This may be caused by a difference in the definition of 'noise' on skeleton features. As shown in Fig. 2, Song et al. [43] assigned 0 to noisy joints, which were defined by the 'missed joints' due to the spatial or temporal occlusions. However, in our experiments, an arbitrary value for joint noise is defined randomly within a bounding box (see Fig. 4). Moreover, the PeGCN not only shows strength on noisy skeletons but also performs comparably to its backbone method on clean skeleton data. This suggests that adopting a better backbone network for the PeGCN could lead to even better performance. Nevertheless, the entire set of experimental results serves to demonstrate the efficiency of PeGCN when applied to skeleton-based action recognition with noisy skeleton samples.

5 Conclusions

In this work, we presented a noise-robust skeleton-based action recognition method based on the graph convolutional networks with predictive encoding for latent space, called a predictively encoded graph convolutional network (PeGCN). In the training step, the PeGCN learns to improve its representation ability of noisy skeleton by

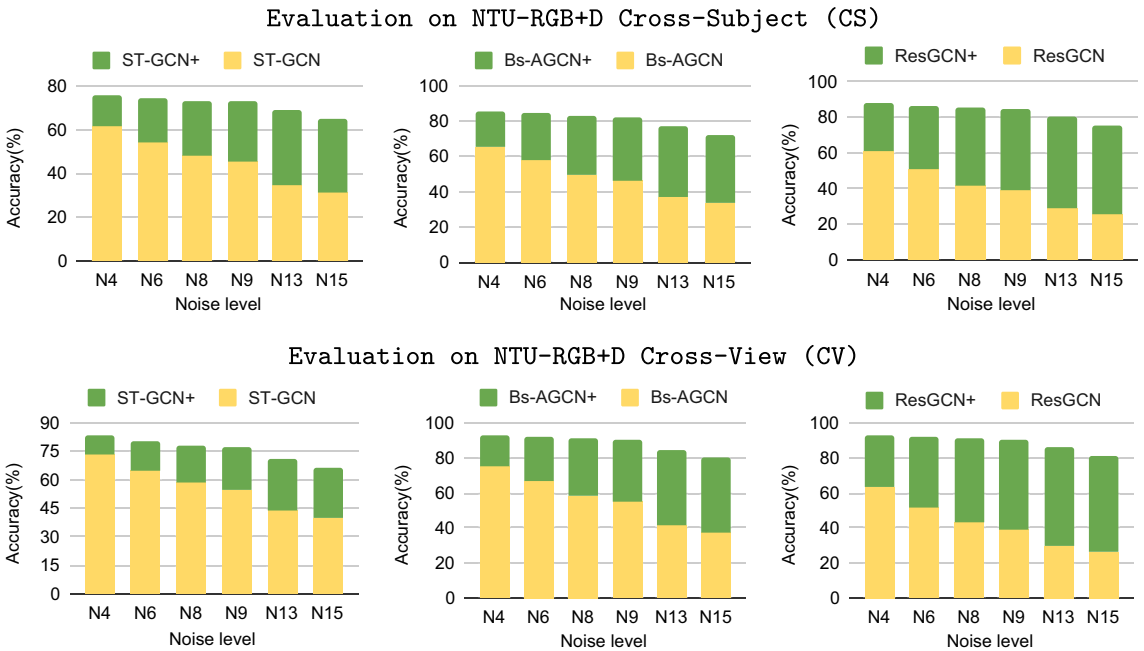

Fig. 7 Illustration of changes in the accuracy for each method before and after applying our noise-robust learning. Symbol + of model name denotes the model to which our predictive encoding learning (PEL) is applied. In each subfigure, the yellow bar and the green bar represent the original accuracy and the improved accuracy with PEL, respectively

predicting complete samples from noisy samples in latent space. The PeGCN increases the flexibility of GCNs and is more suitable for action recognition tasks using skeleton features. When we evaluated the PeGCN on two large-scale action recognition datasets, NTU-RGB+D and Kinetics, it achieved competitive performances to state-of-the-art on both datasets. Although the PeGCN shows considerable action recognition performance on noisy skeletons, training it requires a large numbers of samples. Additionally, the processing speed of the PeGCN is quite slow because it includes two different deep-learning networks.

In future studies, therefore, a unified model will be explored to improve the model's performance on both noisy and clean skeletons, where the current structure consists of two networks connected in sequence. We expect the performance and processing speed to further improve as a result. Although both normal and noisy skeletons were used in this paper for mutual information to extract rich representation, we plan to use only noisy data by applying unsupervised learning without a guide-network. In this way, our method can be applied to any environment even without the availability of a complete skeleton. Lastly, various noise types, not only spatial noise but also temporal noise including partial occlusion, will be studied as well.

Acknowledgments This work was partly supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2014-0-00077, Development of global multitarget tracking and event prediction techniques based on real-time large-scale video analysis) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT). (No. 2019R1A2C208748911).

References

1. Atal BS, Schroeder MR (1970) Adaptive predictive coding of speech signals. Bell Syst Technic J 49(8):1973–1986
2. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Machine Intell 39(12):2481–2495
3. Bengio Y, Senécal JS (2008) Adaptive importance sampling to accelerate training of a neural probabilistic language model. IEEE Trans Neural Netw 19(4):713–722
4. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
5. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y (2018) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. arXiv:181208008
6. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence
7. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:14123555
8. Ding Z, Wang P, Ogunbona PO, Li W (2017) Investigation of different skeleton features for cnn-based 3d action recognition. In: 2017 IEEE International conference on multimedia & expo workshops (ICMEW). IEEE, pp 617–622
9. Du Y, Wang W, Wang L (2015) Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1110–1118
10. Elias P (1955) Predictive coding–i. IRE Trans Inform Theory 1(1):16–24. https://doi.org/10.1109/TIT.1955.1055126
11. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: The IEEE conference on computer vision and pattern recognition (CVPR)
12. Fernando B, Gavves E, Oramas JM, Ghodrati A, Tuytelaars T (2015) Modeling video evolution for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5378–5387
13. Gao T, Packer B, Koller D (2011) A segmentation-aware object detection model with occlusion handling. In: CVPR 2011. IEEE, pp 1361–1368
14. Gao Z, Guo L, Guan W, Liu AA, Ren T, Chen S (2020a) A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. IEEE Trans Image Process 30:767–782
15. Gao Z, Guo L, Ren T, Liu AA, Cheng ZY, Chen S (2020b) Pairwise two-stream convnets for cross-domain action recognition with small data. IEEE Transactions on Neural Networks and Learning Systems
16. Girdhar R, Ramanan D, Gupta A, Sivic J, Russell B (2017) Actionvlad: Learning spatio-temporal aggregation for action classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 971–980
17. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
18. Gutmann M, Hyvärinen A (2010) Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 297–304
19. Hussain M, Chen D, Cheng A, Wei H, Stanley D (2013) Change detection from remotely sensed images: From pixel-based to object-based approaches. In: ISPRS Journal of photogrammetry and remote sensing, vol 80, pp 91–106
20. Jozefowicz R, Vinyals O, Schuster M, Shazeer N, Wu Y (2016) Exploring the limits of language modeling. arXiv:160202410
21. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al. (2017) The kinetics human action video dataset. arXiv:170506950
22. Ke Q, Bennamoun M, An S, Sohel F, Boussaid F (2017) A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3288–3297
23. Kim TS, Reiter A (2017) Interpretable 3d human action analysis with temporal convolutional networks. In: 2017 IEEE Conference on computer vision and pattern recognition workshops (CVPRW). IEEE, pp 1623–1631
24. Li B, Dai Y, Cheng X, Chen H, Lin Y, He M (2017) Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: 2017 IEEE International conference on multimedia & expo workshops (ICMEW). IEEE, pp 601–604
25. Li M, Chen S, Chen X, Zhang Y, Wang Y, Tian Q (2019) Actional-structural graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3595–3603

26. Li YM, Gao Z, Tao YB, Wang LL, Xue YB (2020) 3d object retrieval based on non-local graph neural networks. Multimed Tools Appl 79(45):34011–34027
27. Liu J, Shahroudy A, Xu D, Wang G (2016) Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision. Springer, pp 816–833
28. Liu J, Wang G, Duan LY, Abdiyeva K, Kot AC (2017a) Skeleton-based human action recognition with global context-aware attention lstm networks. IEEE Trans Image Process 27(4):1586–1599
29. Liu J, Rahmani H, Akhtar N, Mian A (2019) Learning human pose models from synthesized data for robust rgb-d action recognition. Int J Comput Vis 127(10):1545–1564
30. Liu M, Liu H, Chen C (2017b) Enhanced skeleton visualization for view invariant human action recognition. Pattern Recogn 68:346–362
31. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:13013781
32. Mnih A, Teh YW (2012) A fast and simple algorithm for training neural probabilistic language models. arXiv:12066426
33. Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv:180703748
34. Peng W, Hong X, Chen H, Zhao G (2019) Learning graph convolutional network for skeleton-based human action recognition by neural searching. arXiv:191104131
35. Qian R, Meng T, Gong B, Yang MH, Wang H, Belongie S, Cui Y (2020) Spatiotemporal contrastive video representation learning. arXiv:200803800
36. Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019
37. Shi L, Zhang Y, Cheng J, Lu H (2019a) Skeleton-based action recognition with directed graph neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7912–7921
38. Shi L, Zhang Y, Cheng J, LU H (2019b) Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. arXiv:191206971
39. Shi L, Zhang Y, Cheng J, Lu H (2019c) Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12026–12035
40. Si C, Jing Y, Wang W, Wang L, Tan T (2018) Skeleton-based action recognition with spatial reasoning and temporal stack learning. In: Proceedings of the european conference on computer vision (ECCV), pp 103–118
41. Si C, Chen W, Wang W, Wang L, Tan T (2019) An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1227–1236
42. Song S, Lan C, Xing J, Zeng W, Liu J (2017) An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Thirty-first AAAI conference on artificial intelligence
43. Song YF, Zhang Z, Wang L (2019) Richly activated graph convolutional network for action recognition with incomplete skeletons. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 1–5
44. Song YF, Zhang Z, Shan C, Wang L (2020) Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In: Proceedings of the 28th ACM international conference on multimedia (ACMMM), association for computing machinery, New York, NY, USA, pp 1625–1633. https://doi.org/10.1145/3394171.3413802
45. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: The IEEE conference on computer vision and pattern recognition (CVPR)
46. Tang Y, Tian Y, Lu J, Li P, Zhou J (2018) Deep progressive reinforcement learning for skeleton-based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5323–5332
47. Thakkar K, Narayanan P (2018) Part-based graph convolutional network for action recognition. arXiv:180904983
48. Tran C, Trivedi MM (2011) 3-d posture and gesture recognition for interactivity in smart spaces. IEEE Trans Indust Inform 8(1):178–187
49. Vondrick C, Pirsiavash H, Torralba A (2016) Anticipating visual representations from unlabeled video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 98–106
50. Wang C, Xu D, Zhu Y, Martín-Martín R, Lu C, Fei-Fei L, Savarese S (2019a) Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3343–3352
51. Wang L, Koniusz P, Huynh DQ (2019b) Hallucinating idt descriptors and i3d optical flow features for action recognition with cnns. In: Proceedings of the IEEE international conference on computer vision, pp 8698–8708
52. Wang X, Han TX, Yan S (2009) An hog-lbp human detector with partial occlusion handling. In: 2009 IEEE 12th international conference on computer vision. IEEE, pp 32–39
53. Wang X, Gao L, Wang P, Sun X, Liu X (2017a) Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length. IEEE Trans Multimed 20(3):634–644
54. Wang X, Shrivastava A, Gupta A (2017b) A-fast-rcnn: Hard positive generation via adversary for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2606–2615
55. Xia L, Chen CC, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, pp 20–27
56. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence
57. Yu J, Yow KC, Jeon M (2018) Joint representation learning of appearance and motion for abnormal event detection. Mach Vis Appl 29(7):1157–1170
58. Yu J, Park S, Lee S, Jeon M (2019) Driver drowsiness detection using condition-adaptive representation learning framework. IEEE Trans Intell Transp Syst 20(11):4206–4218. https://doi.org/10.1109/TITS.2018.2883823
59. Zhang P, Lan C, Xing J, Zeng W, Xue J, Zheng N (2017) 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
60. Zhang Z (2012) Microsoft kinect sensor and its effect. IEEE Multimed 19(2):4–10

Yongsang Yoon received a B.S. degree in computer science from Chonnam National University, Gwangju, South Korea, in 2015. He is currently pursuing a Ph.D. degree with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju. His research interests include artificial intelligence, machine learning, and pattern recognition.

Jongmin Yu received a B.S. degree in computer science from Chungnam National University, Daejeon, Republic of Korea, in 2013. He is a joint PhD candidate at the School of Electrical Engineering and Computer Science in Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, and the School of Electrical Engineering, Computing and Mathematical Sciences in Curtin University, Perth, Western Australia, Australia. His current research interests include artificial intelligence, machine learning, and pattern recognition.

Moongu Jeon received a B.S. degree in architectural engineering from Korea University, Seoul, Korea, in 1988 and earned M.S. and Ph.D. degrees in computer science and scientific computation from the University of Minnesota, Minneapolis, MN, USA, in 1999 and 2001, respectively. As a postgraduate researcher, he worked on optimal control problems at the University of California at Santa Barbara, Santa Barbara, CA, USA, from 2001 to 2003 and then moved to the National Research Council of Canada, where he worked on sparse representation of high-dimensional data and image processing until July 2005. In 2005, he joined the Gwangju Institute of Science and Technology, Gwangju, Korea, where he is currently a full professor in the School of Electrical Engineering and Computer Science. His current research interests lie in machine learning, computer vision and artificial intelligence.