applied sciences

MDPI

Article

A Deep Attention Model for Action Recognition from Skeleton Data

Yanbo Gao [1,2], Chuankun Li [3,*], Shuai Li [4], Xun Cai [1,2], Mao Ye [5] and Hui Yuan [4]

1 School of Software, Shandong University, Jinan 250101, China; ybgao@sdu.edu.cn (Y.G.); caixunzh@sdu.edu.cn (X.C.)
2 Shandong University-Weihai Research Institute of Industrial Technology, Weihai 264209, China
3 State Key Laboratory of Dynamic Testing Technology, School of Information and Communication Engineering, North University of China, Taiyuan 030051, China
4 School of Control Science and Engineering, Shandong University, Jinan 250100, China; shuaili@sdu.edu.cn (S.L.); huiyuan@sdu.edu.cn (H.Y.)
5 School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; cvlab.uestc@gmail.com
* Correspondence: chuankun@nuc.edu.cn

Abstract: This paper presents a new IndRNN-based deep attention model, termed DA-IndRNN, for skeleton-based action recognition to effectively model the fact that different joints are usually of different degrees of importance to different action categories. The model consists of (a) a deep IndRNN as the main classification network to overcome the limitation of a shallow RNN network in order to obtain deeper and longer features, and (b) a deep attention network with multiple fully connected layers to estimate reliable attention weights. To train the DA-IndRNN, a new triplet loss function is proposed to guide the learning of the attention among different action categories. Specifically, this triplet loss enforces intra-class attention distances to be smaller than inter-class attention distances and at the same time to allow multiple attention weight patterns to exist for the same class. The proposed DA-IndRNN can be trained end-to-end. Experiments on the widely used datasets, including the NTU RGB + D dataset and UOW Large-Scale Combined (LSC) Dataset, have demonstrated that the proposed method can achieve better and stable performance than the state-of-the-art attention models.

Keywords: skeleton-based action recognition; IndRNN; RNN; attention model

Citation: Gao, Y.; Li, C.; Li, S.; Cai, X.; Ye, M.; Yuan, H. A Deep Attention Model for Action Recognition from Skeleton Data. *Appl. Sci.* **2022**, *12*, 2006. https://doi.org/10.3390/app12042006

Academic Editor: João M. F. Rodrigues

Received: 26 January 2022
Accepted: 10 February 2022
Published: 15 February 2022

1. Introduction

Human action recognition has received increasing interest in the past due to its wide range of applications in video analytics, robotics, health monitoring, and autonomous driving. The success of deep learning in computer vision has driven the development of many deep models [1–9] for action recognition. Among these models, recurrent neural network (RNN) [10–16] is one of the popular ones because of its capability of modeling sequential data. Recently, RNNs are further augmented with attention models [17,18] to explicitly model the observation that discriminative information presents in different body parts at different time steps. Noticeable improvement in performance has been attained [19,20].

This paper is concerned with two fundamental and challenging issues in an attention-based RNN for action recognition from skeleton data, where the attention weights are associated with joints of the skeletons. First, the state-of-the-art attention models, such as those presented in [19,20], lack proper regularization on attention weights. The attention weights would be sufficiently different for different actions, and they can be similar with small differences for same actions to accommodate different performing styles. For example, the joints of legs in action "kicking" would have higher attention weights than other joints,

as would the joints of arms in action "boxing". Therefore, the attention on joints for different actions is different, that is, the attention weights between two skeleton samples of "kicking" should be more similar than the attention weights between a skeleton sample of "kicking" and a skeleton sample of "boxing". This makes it possible to regularize attention weights for different action categories. In addition, multiple sets of joints may be discriminative for different samples of a same class of actions. For example, one subject may perform the "hand waving" with their left hand while another one may perform it with their right hand. Therefore, while similarities exist in the attention weights for one action class, there may also be differences, which should also be considered in the attention regularization in order to achieve robust recognition.

Second, the general principle that the deeper the network, the better in extracting discriminative features is hardly implementable using a conventional RNN, such as the Vanilla RNN and long short-term memory (LSTM), due to the notorious gradient vanishing and exploding problems [21]. Attention model-based RNNs for action recognition usually only include one or two fully connected layers to obtain an attention model and one or two LSTM layers for the classification as in [19,20]. Such shallow networks are hardly able to explore the long-range dependency both temporally and spatially and a deep (e.g., multiple layers) RNN is expected to improve the performance as observed in [14]. In addition, one fully connected layer in estimating the attention weight tends to trap the end-to-end training to a local optimum as shown in the experiments. Such a local optimum issue cannot be resolved by the double stochastic attention regularization [18], which aims to encourage the model to pay equal attention to every joint over a sequence of skeletons.

To address these two issues, this paper proposes the following:

- A new deep attention model in which the IndRNN model [14] is adopted to build up a deep RNN for classification, and multiple fully connected layers are employed to estimate the attention weights for each joint at each time step. An ablation study has shown that the proposed deep attention architecture provides much more stable and better performance than the shallow counterparts.
- A new triplet loss function to regulate the attention among different action categories. This triplet loss function is further extended with a sample to class distance to enforce the intra-class attention distances to be no larger than the inter-class distances and at the same time to allow different sets of attention weights within the same class.

Experimental results have shown that the proposed deep attention architecture and the new loss function improves significantly the performance of classification and that the attention model learned is much more stable compared with the traditional attention models [19,20]. In addition, the double stochastic attention regularization [18] is no longer required to train the network.

The rest of the paper is organized as follows. The related work on the skeleton-based action recognition and the attention models are reviewed in Section 2. The proposed model is explained in Section 3, and the experimental results and analysis are presented in Section 4. Finally, the conclusion is drawn in Section 5.

2. Background

A large number of skeleton-based action recognition methods have been proposed in the literature. Among them, many methods employ the deep learning models, including both the convolutional neural networks (CNN) and the recurrent neural networks (RNN). CNN-based methods [22–26] usually summarizes the information from all frames into one image and then apply CNN for classification on this single image. On the contrary, RNN-based methods [19,20,27–31] sequentially process the frames and classify the sequences after all the frames are given. For the CNN-based methods, in [32], joint distance maps are developed to capture the spatial–temporal information. In [33], skeleton optical spectra images are designed to map the skeleton sequence into a single image. Similar ideas of mapping the sequence information into one frame have also been explored in [34,35].

Since the method proposed in this paper is an RNN-based model, we mostly discuss the related RNN-based methods here. Most of the existing RNN-based methods employ the long short-term memory network (LSTM), which can better maintain the long-term memory. On top of the original LSTM models, many models have been developed to take advantage of the specific features of skeletons. Considering that body joints move together in groups, in [27], a hierarchical recurrent neural network was proposed where different parts of the body are first processed with different RNNs and then concatenated together for the whole body action. Similarly, in [29], a part-aware LSTM model was proposed where the LSTM cell is split into different parts for different groups of joints of the body in order to explore the joint groups. Since some skeleton data may be noisy, a trust gate is further added in the LSTM model in [30]. In [36], the co-occurrence of joints is explored by adding a fully connected layer before LSTM to learn joint connections. There are also methods exploring other types of features in the skeleton data. In [31], instead of processing the joint coordinates, a geometric feature was proposed for skeletons to explore the geometric relationships between different joints. In [28], a differential gating scheme was proposed for LSTM, which emphasizes the change in information gain caused by the salient motions between different frames. These methods consider and process all joints equally at all time steps for all actions, which is against the intuition that different joints may contribute differently to the classification of actions.

RNN-based methods are recently augmented by the incorporation of attention models [17,18,37,38] to explicitly model the observation that, for different actions, different joints may show different degrees of importance in classification. In [19], an attention model was proposed to assign different weights to different joints at different time steps. An additional temporal weight is assigned to the features obtained at different time steps for the final classification. Since no ground truth is available for the attention model, the attention weights are often treated as latent variables and trained by the final classification objective. The doubly stochastic attention regularization [18] is most widely used to encourage the model to pay equal attention to every joint over the sequence in order to avoid attention weights only being assigned to one or two joints. In [20], the attention model was also used where the global context over the whole sequence is employed to obtain the attention weights. These attention models are similar to those used in the image-based action recognition [18], where no direct loss is applied to regulate the attention weights other than the doubly stochastic attention regularization, and they often fail to meet the requirement on similar and multiple intra-class attention weights and different inter-class attention weights.

The most recent independent recurrent neural network (IndRNN) [14,39] provides an effective solution to the gradient vanishing and exploding problem in training a multiple-layer RNN, which allows deep networks to be constructed and to learn long-term dependency. Specifically, preliminary experiments using multiple layers of the basic IndRNN have shown that better performance than LSTM-based networks on the skeleton-based action recognition can be attained. Therefore, IndRNNs are adopted in this paper to construct a deep attention network for action recognition, and a new regularization is developed to train the network. In addition to RNNs, there are also methods exploring temporal convolution and graph convolution for temporal processing such as [40,41], which are not further discussed here.

3. Proposed Method

3.1. IndRNN-Based Deep Attention Model

In this paper, the independently recurrent neural network (IndRNN) [14] is used as a basic RNN component to construct a deep RNN for classification to leverage IndRNN's capability of learning deeper and longer features than LSTM. It follows

$$\mathbf{h}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{u} \odot \mathbf{h}_{t-1} + \mathbf{b}) \tag{1}$$

where $\mathbf{x}_t \in \mathbb{R}^M$ and $\mathbf{h}_t \in \mathbb{R}^N$ are the input and hidden state at time step t, respectively. $\mathbf{W} \in \mathbb{R}^{N \times M}$, $\mathbf{u} \in \mathbb{R}^N$, and $\mathbf{b} \in \mathbb{R}^N$ are the weight for the current input, the weight for the recurrent input, and the bias of the neurons. $\odot$ represents the Hadamard product (element-wise multiplication). σ is an element-wise activation function of the neurons, which is the ReLU (rectified linear unit) in this paper, and N is the number of neurons in this IndRNN layer. Each neuron in one layer is independent from the others, and connection between neurons is achieved by stacking two or more layers of IndRNNs as presented later.

Figure 1a shows the framework of the IndRNN-based deep attention model for skeleton-based action recognition. It consists of a main classification network and an attention network. The main classification network is composed of several IndRNN layers, and batch normalization layers are inserted after each IndRNN layer (ignored in Figure 1a for simplicity). Residual connections are also used to further facilitate the gradient propagation across layers, and each residual block consists of two IndRNN layers. A fully connected layer (FC layer) is added at the final time step for classification (also ignored in Figure 1a for simplicity). Due to the use of hidden states at each time step to obtain the attention model, the statistics for the batch normalization is estimated for each time step, while the parameters for the affine mapping is shared over time.

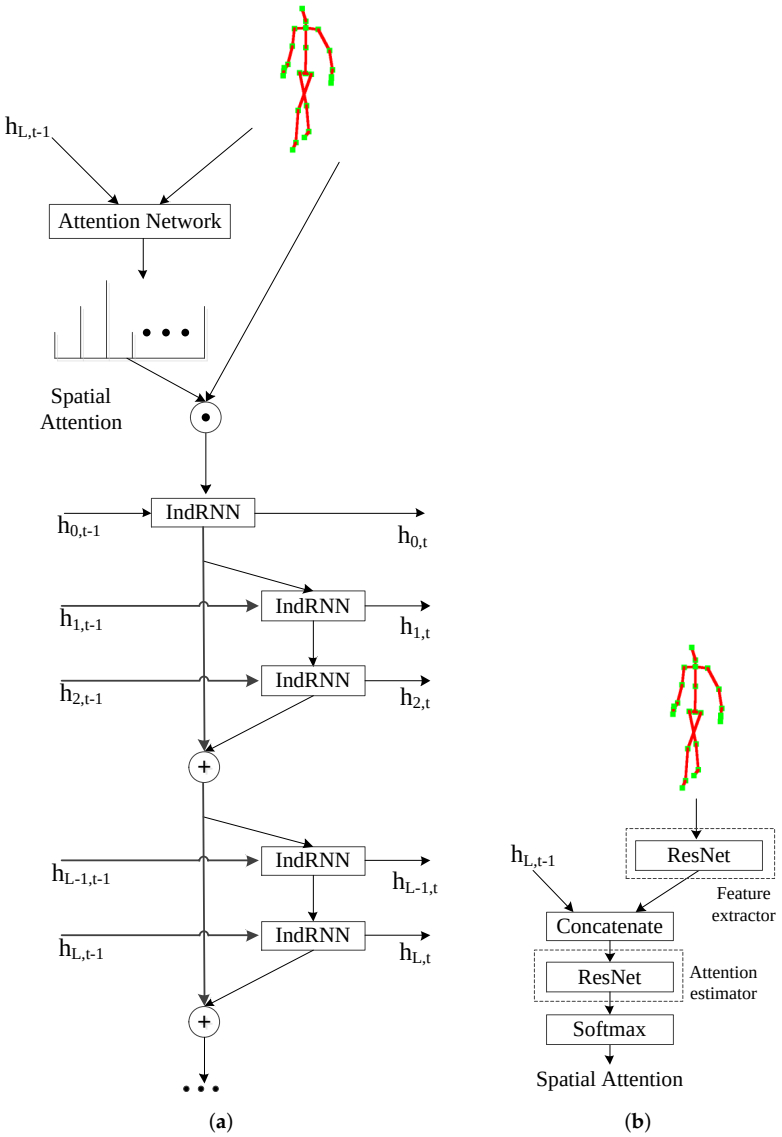


Figure 1. The framework of the spatial IndRNN-based attention models for skeleton-based action recognition. (**a**) The attention network obtains the attention based on the last hidded state and the current input. The main classification network is a deep residual IndRNN network where each residual block contains two IndRNN layers. (**b**) Illustration of the attention network used to process the input and the last hidden state from IndRNN.

The attention for each time step (s_t) is obtained using an attention network based on the current input and the hidden state of the last IndRNN layer at the previous time step. Unlike the conventional attention network that consists of only one FC layer, a multiple-layer network is proposed as the attention network. It avoids the issue that the conventional one-layer attention network cannot robustly estimate attention, leading to the attention being mainly placed on joints with larger movements. Specifically, since the hidden state of the last IndRNN layer captures the high-level information after going through several IndRNN layers, a few FC layers are first used to extract efficient features of the input skeleton before concatenating it with the hidden state of IndRNN. Figure 1b illustrates the attention network, where residual networks are used with residual connections across every two FC layers. The residual network used to extract the features of the input skeleton, referred to as feature extractor, consists of 5 FC layers, while the residual network used to process the concatenated features, referred to as attention estimator, consists of 4 FC layers.

Details on the architecture will be further explained in Section 4. A softmax function over all the joints is added at the end to produce the attention weights.

The input to the IndRNN net is $x'_t = (x'_{t,1}, x'_{t,2}, \ldots, x'_{t,K})$ with $x'_{t,i} = s_{t,i} \cdot x_{t,i}$ where $x_{t,i}$ and $s_{t,i}$ represent the feature and attention weight of the i-th joint. Cross-entropy loss is used as the classification objective. Conventionally, the doubly stochastic attention regularization [18,19] would be used to regularize the attention weights. However, the proposed deep attention model can be trained robustly without this regularization term and thus is not employed in the training probably because of the deeper networks for classification and attention estimation. To address the first issue discussed in the introduction, a new loss function, referred to as *Triplet Attention Loss*, is developed to guide the learning of attention weights.

3.2. Triplet Attention Loss

It is known that different joints can be of different degrees of importance to different action classes. Therefore, the attention weights for different joints are different for different action classes. For example, "Kicking" mostly focuses on the leg, while "Punching" mostly focuses on the hand. Accordingly, the attention weights assigned to the informative joints are supposed to be larger than others, and the attention weights for samples from one class are supposed to be more similar than those from different classes. However, this cannot be achieved in the current attention model framework learned completely by the final objective without any direct constraints on the attention model.

To address the above problem, a new triplet loss function is proposed to guide the learning of the attention weights in addition to the final classification objective. This new loss function enforces the intra-class attention distance to be smaller than the inter-class attention distance. As mentioned before, for skeleton-based attention models, the attention weights are assigned to the joints and the mapping between attention and joints is fixed. Therefore, for different samples in the same class and different classes, the attention distance over the joints can be calculated, and thus, the triplet loss can be implemented. The details on the new triplet loss functions are defined as follows.

For a skeleton sample v_i (*anchor*), let the attention weights on the sample at time step t be $s^a_{i,t}$. A sample from the same action class represents the *positive* sample whose attention weights is $s^p_{i,t}$, and one sample from different action classes represents the *negative* sample whose attention weights by $s^n_{i,t}$. The following constraint stands.

$$||s^a_{i,t} - s^p_{i,t}||^2_2 + \alpha < ||s^a_{i,t} - s^n_{i,t}||^2_2 \tag{2}$$

where α is a margin that is enforced between positive and negative pairs. This constrains the intra-class attention distance to be smaller than the inter-class attention distance by at least α.

Accordingly, a triplet loss ($L_{tri,t}$) [42,43] for the spatial attention model in each frame can be defined as

$$L_{tri,t} = \sum_{i}^{N} [||s^a_{i,t} - s^p_{i,t}||^2_2 - ||s^a_{i,t} - s^n_{i,t}||^2_2 + \alpha]_+ \tag{3}$$

where N is the number of samples. When the triplet loss function is used to obtain embeddings for different classes [42,43], α is set to a positive value to avoid the trivial solution that embeddings for different classes are the same. However, in the attention models for skeleton-based action recognition, the attention weights are guided by the class classification objective in addition to the above triplet loss. Therefore, α can be set to zero, which only encourages the distance between the attention weights of the anchor and the positive sample to be no larger than that between the anchor and the negative sample. This allows different action classes to share similar attention weights.

On the other hand, since multiple sets of attention weights may exist for one action, the above triplet loss function may reduce the number of plausible attention weights for

one action. To overcome this issue, a new triplet loss is proposed based on a sample to class distance, instead of the distance between the attention weights of anchor and one positive.

Figure 2 illustrates the new triplet loss. The anchor to the positive class distance is defined as the minimum distance between the anchor and multiple positive samples, $\min_{m=1,2,...,M} ||s_{i,t}^{a} - s_{i,t}^{p_m}||_2^2$, where M is the number of samples used in the positive class. Through learning, the attention of the anchor sample is pushed to be closer to the attention of the positive classes. In this way, the attention weights for the anchor only need to be closer to any sample in the positive class than the negative sample, which allows for multiple sets of attention weights for each action. This also compensates, to some extent, the differences of attention (if there is any) due to the styles of performing the actions or other factors.

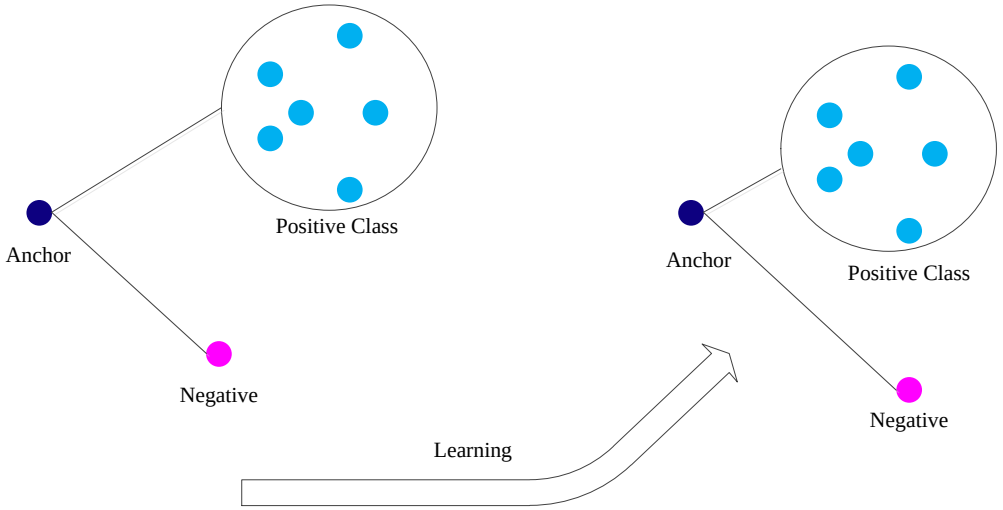


Figure 2. Illustration of the new triplet loss with sample to class distance.

To reduce the computation, the positive class is represented by a few randomly selected positive samples. Accordingly, the new triplet loss ($L_{tri,t}$) follows

$$L_{tri,t} = \sum_{i}^{N} [\min_{m=1,2,...,M} ||s_{i,t}^{a} - s_{i,t}^{p_m}||_2^2 - ||s_{i,t}^{a} - s_{i,t}^{n}||_2^2 + \alpha]_+. \quad (4)$$

Assuming the probability of the attention set used by the current anchor sample is p_a, the probability of at least one sample in the selected positive class samples sharing the attention set as the current anchor sample is $1 - (1 - p_a)^M$. Comparing with the probability (p_a) of drawing one random positive sample sharing the attention set, the probability is higher with $M > 1$, and it is much higher when M is large.

Notice that different action samples may be not well aligned in the time domain in practice because of the varying starting points and speeds of an action being performed. Accordingly, the frame-by-frame distance between the anchor sample and the positive sample could be large even if they may share similar attention weights on some key frames. This issue has been widely studied in the literature and could be addressed with preprocessing the sequences by time dynamic warping if the increased computation complexity is affordable. In most of the training datasets for action recognition, each video sample is a short video clip containing a single action. In the training process, each video is first divided to T sub-sequences with the same length, and one frame is randomly selected from each sub-sequence [29]. The attention weights can be averaged over a few frames to make each segment roughly aligned. Since the triplet loss function is only used in training, this processing does not affect the testing.

In all, the final objective function used for training is as follows.

$$L = L_c + \lambda_2 \sum_{t=0}^{T} L_{tri,t} \tag{5}$$

where L_c represents the class classification loss using the typical cross-entropy loss ($-\sum_{i=1}^{C} y_i \log \hat{y}_i$, where y_i and $\hat{y}_i$ are the groundtruth label and the predicted label, respectively). $\sum_{t=0}^{T} L_{tri,t}$ represents the triplet loss on the attention model over time where T is the length of the sequence.

4. Experimental Results

The proposed deep attention model has been evaluated in two widely used datasets, i.e., the NTU RGB + D dataset [29] and UOW Large-Scale Combined (LSC) Dataset [44], which covers a wide range of actions.

4.1. Results on NTU RGB + D Dataset

The NTU RGB + D dataset [29] is a large available action recognition dataset with skeletons. It contains 56,880 sequences (over four million frames) of 60 action classes, including one-person daily activities and two-person interactions. It was collected by three Kinect v2 cameras with 17 different setups, and 40 subjects are involved in performing the actions. Two evaluation protocols are suggested for this dataset including Cross-Subject (CS) and Cross-View (CV) settings. In the training, 5% of the training data was reserved as evaluation data, as suggested in [29]. Two skeletons (25 joints per skeleton) were used as input, and if only one is present in the sample, the second was set as zero. For this dataset, when multiple skeletons are present in the scene, the skeleton number captured by Kinect may be changed over time, especially when the number of the skeletons is changed. Therefore, an alignment process, by comparing the distance of all the joints between different skeletons is first applied to keep the same number assigned to the skeleton of the same subject. This is only performed once, as preprocessing and the processed skeleton data were used for training and testing the network. For both training and testing, each sequence was first divided to 20 segments of the same length, and one frame is randomly selected from each segment [30].

The hyperparameters of the IndRNN-based deep attention model used in the experiments are as follows. For the main IndRNN classification network, seven IndRNN layers with residual connections are used, and each layer contains 512 neurons. Five and four layers are used for extracting the features of the input skeleton and processing the concatenated features in the attention network, respectively. The joint coordinates of two persons (of dimension $25 * 2 * 3$) are used as input. The training setup is similar as in [14], where the batch size was set to 128. The Adam optimization is used with the initial learning rate 2×10^{-4} and decayed by 10 once the evaluation accuracy does not increase (with patience 20). Dropout is applied after each layer with a dropping probability of 0.45 and 0.3 (larger than [14]) for the CS and CV settings, respectively.

The proposed triplet loss function was evaluated using the above deep attention network. The overall performance of the proposed model in comparison with the existing methods is shown in Table 1, where the proposed IndRNN-based deep attention model is denoted by DA-IndRNN. It can be seen that the proposed model significantly outperformed the traditional RNN-based methods such as [29–31] and also the very recent attention model-based methods such as [19,20]. Moreover, it can be observed that compared with the traditional LSTM-based attention model [19], the performance improvement under the CS (cross-subject) setting is almost by 10 percentage points and it is higher than 7.5 percentage points under the CV (cross-view) setting. This is mostly because the performing styles are more diverse among different subjects than among different views of the same subject, and the proposed model allows different styles with different attention weight patterns.

Note that since this paper focuses on the investigation of the RNN-based skeleton action recognition, the graph convolution-based ones [40,41] are not discussed and compared.

Table 1. Results of the skeleton-based action recognition on NTU RGB + D dataset in comparison with the existing methods.

Method	CS	CV
Deep learning on Lie Group [45]	61.37%	66.95%
JTM + CNN [22]	73.40%	75.20%
Res-TCN [24]	74.30%	83.10%
SkeletonNet (CNN) [23]	75.94%	81.16%
JDM + CNN [32]	76.20%	82.30%
Clips + CNN + MTLN [25]	79.57%	84.83%
Enhanced Visualization + CNN [4]	80.03%	87.21%
ST-GCNN [46]	81.5%	88.3%
1 Layer RNN [29]	56.02%	60.24%
2 Layer RNN [29]	56.29%	64.09%
1 Layer LSTM [29]	59.14%	66.81%
2 Layer LSTM [29]	60.09%	67.29%
1 Layer PLSTM [29]	62.05%	69.40%
2 Layer PLSTM [29]	62.93%	70.27%
JL_d + RNN [31]	70.26%	82.39%
ST-LSTM + Trust Gate [30]	69.20%	77.70%
GCA-LSTM [20]	74.4%	82.8%
STA-LSTM [19]	73.40%	81.20%
IndRNN [14]	81.80%	87.97%
Proposed DA-IndRNN	83.24%	88.70%

4.1.1. Evaluation of the Deep Attention Network

It is known that the skeleton data can be noisy. Therefore, in addition to the hidden state of the previous time step, the current input information is also used as input to obtain the attention, as shown in Section 3.1. Moreover, we show that it is important to explore relatively deep features to provide robust performance. Figure 3 shows comparisons among different configurations of the feature extractor and attention estimator for the attention network. Each figure shows the performance over 10 training processes of the corresponding configuration of the attention network. It can be clearly seen that with only a one-FC-layer attention estimator and without a feature extractor, the training process is likely to be trapped to a local optimum with poor performance, as shown in Figure 3a. As the number of layers in the attention estimator and the feature extractor increases, the training process becomes stable and less likely to be trapped to a local optimum. For a network with a four-FC-layer attention estimator and a five-FC-layer feature extractor, the training converges almost monotonically with improving performance.

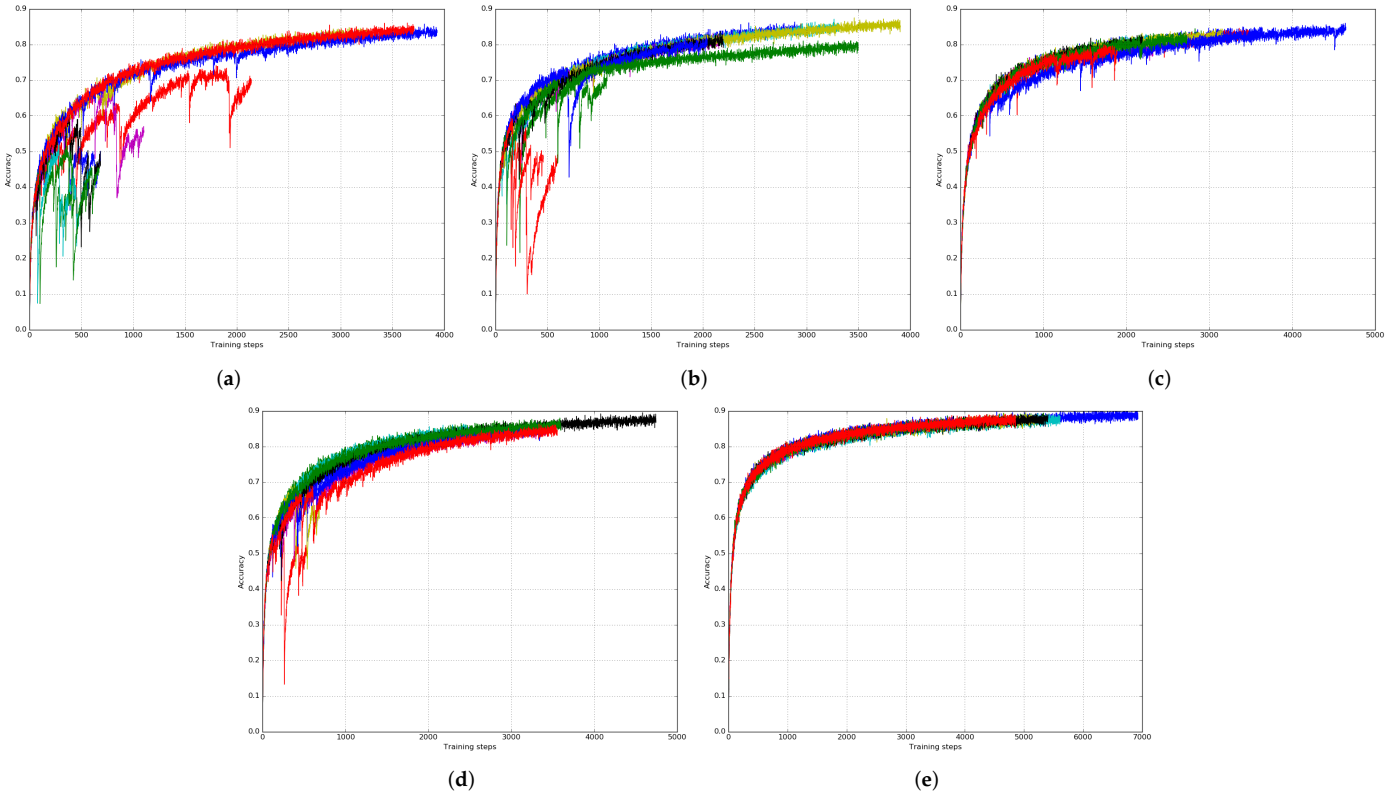


Figure 3. Performance comparison of different attention models obtained with 10 training processes on the CS setting of the NTU RGB + D dataset [29]. (**a**–**d**) represent the attention model using a one-FC-layer attention estimator with no feature extractor, a one-FC-layer attention estimator with a one-FC-layer feature extractor, a two-FC-layer attention estimator with no feature extractor, and a two-FC-layer attention estimator with a one-FC-layer feature extractor, respectively. (**e**) represents the proposed deep attention model using a four-FC-layer attention estimator with a five-FC-layer feature extractor.

4.1.2. Comparison of the DA-IndRNN against the Shallow LSTM-Based Attention Model

To demonstrate the advantage of the proposed model against a shallow LSTM-based attention model [19], the difference between the normalized confusion matrices (%) of the proposed model minus that of the shallow LSTM-based attention model is shown in Figure 4a. Since the number of classes (60) is too large to show the entire difference matrix, Figure 4b only shows part of the difference confusion matrix. The 10 rows represent the classes with the 10 largest differences between the proposed model and the shallow LSTM-based attention model. The columns represent the classes that are confused with one of the classes in the rows, and the difference between the proposed model and the shallow LSTM-based attention model is at least 2%. The positive values of the main highlighted diagonal elements (where the true label equals the predicted label) represent the improvement in percentage points achieved by the proposed model compared to the shallow model for these classes. The negative values of the other elements represent that error in percentage points reduced by the proposed model. From the figure, it can be seen that the proposed model significantly improves the performance of these classes with up to 29 percentage points, and it reduces the confusion among classes, especially for classes performed with same joints but small differences, such as Class 13 ("teat up paper") and Class 11 ("reading"), or Class 31 ("pointing to something with finger") and Class 32 ("taking a selfie"), both involving the movement of the hands.

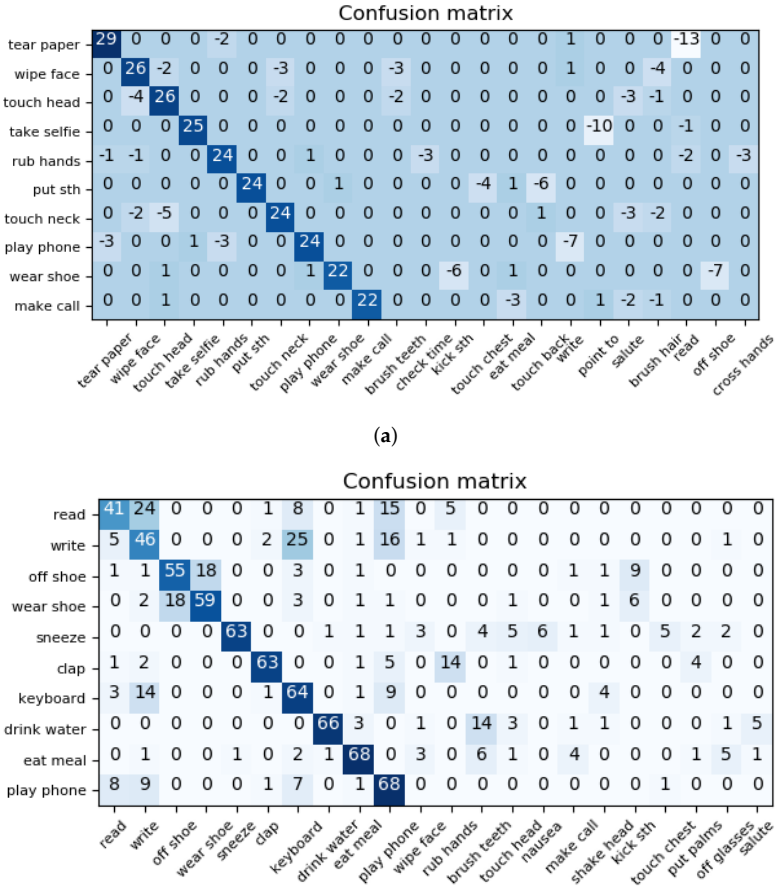
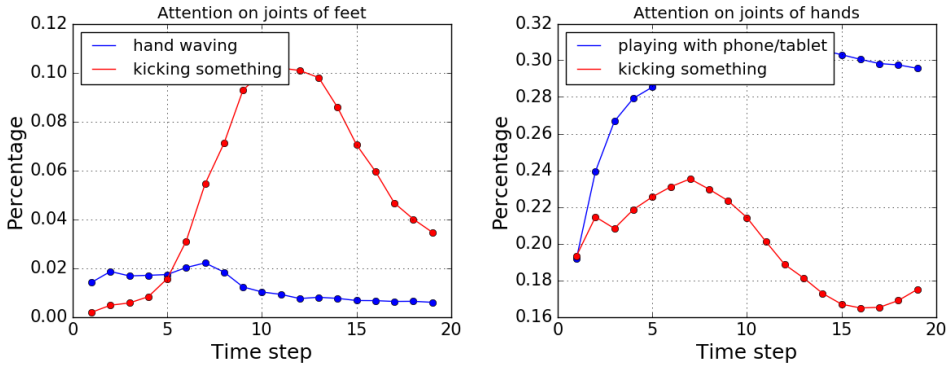


Figure 4. Illustration of the normalized confusion matrices (%) for the proposed model and its difference with the shallow LSTM-based attention model on the NTU dataset. Note that there are 60 classes in total, which is too large to show the entire difference matrix. Therefore, only part of the difference confusion matrix is shown. For the difference of the confusion matrix (confuse matrix of the proposed model minus that of the shallow LSTM-based attention model), the 10 rows represent the classes with the 10 largest differences between the proposed model and the shallow LSTM based attention model. The columns represent the classes that are confused with one of the classes in the rows and the difference between the proposed model and the shallow LSTM-based attention model is at least 2%. For the confusion matrix of the proposed model, the top 10 most confused classes are shown in the 10 rows. The columns represent the classes that are confused with one of the classes in the rows by at least 2%. Note that for the purpose of display, the action names are simplified without losing their meaning. Readers are referred to [29] for the full action names. (**a**) Difference of the confusion matrix between the proposed model and the shallow LSTM based attention model. (**b**) Confusion matrix of the proposed model.

To further show the detailed performance of the proposed model, Figure 4b shows part of the confusion matrix of the proposed model. The top 10 most confused classes are shown in the 10 rows. The columns represent the classes that are confused with one of the classes in the rows by at least 2% . The values of the main highlighted diagonal elements represent the accuracy of the proposed model, and the larger the better. The values of other elements represent the errors, and the smaller the better. Despite the performance improvement of the proposed model, it seems that the proposed model still suffers from distinguishing the order of each movement. Taking Class 16 ("wear a shoe") and Class 17 ("take off a shoe") for example, these two classes with similar movements but different orders cannot be classified well by the proposed model. Therefore, RNN models with explicit considering of the orders may be needed, which will be investigated in the future.

4.1.3. Verification of the Learned Attention

Figure 5 shows the attention on the joints of hands and feet over time for different classes. The feet are represented by joints 15, 16, 19, and 20 (left ankle, left foot, right ankle, and right foot), and hands include joints 7, 8, 22, 23, 11, 12, 24, and 25 (left wrist, left hand, tip of the left hand, left thumb, right wrist, right hand, tip of the right hand, and right thumb). The average percentages of the joints being actively focused (where the joint attention weight is larger than 0.04 which represents the case that all 25 joints of a skeleton are equally weighted if the second person is not attended at all) are shown. Figure 5a shows the attention on the joints of feet over time for action "kicking something" and action "hand waving". It can be seen that for "kicking something", with the progress of the actions in time, more joints of feet are being focused, until it reaches the midpoint of the action, and toward the end of the action, it starts to shift back to the original state. On the other hand, for "hand waving", the attention on the joints of feet decreases over time. This agrees with the intuition that "kicking something" focuses on the joints of feet while "hand waving" does not. Figure 5b shows the attention on the joints of hands over time for action "playing with phone/tablet" and action "kicking something". Similar behavior as Figure 5a can be observed where attention on the joints of the hands improves for "playing with phone/tablet".



(**a**) (**b**)

Figure 5. Compasion of the attention on the joints of feet and hands over time for different action classes. (**a**) Progress of the attention on the joints of feet over time for "hand waving" and "kicking something". (**b**) Progress of the attention on the joints of hands over time for "playing with phone/tablet" and "kicking something".

4.2. Results on UOW Large-Scale Combined (LSC) Dataset

The UOW LSC dataset [44] is a large dataset composed of nine publicly available datasets including MSR Action3D Ext [47], UTKinect [48], MSR DailyActivity [49], MSR ActionPair [50], CAD120 [51], CAD60 [52], G3D [53], RGBD-HuDa [54], and UTD-MHAD [55]. There are 94 actions, from 107 subjects and 4953 samples in total. However, some samples in some action classes do no contain skeleton modality and thus are excluded in the experiment, resulting in 88 action classes and 3897 samples. Due to the nature of the LSC dataset combined from several existing datasets, the intra-class variation is much larger than a single dataset captured in a rather fixed environment. Moreover, the number of samples in different action classes are highly imbalanced, ranging from over 100 samples to only four samples in a class. The protocols developed in [44] were used for evaluation, including random cross-subject and random cross-sample. The average recall and precision of all classes as suggested in [44] is used for comparison. Table 2 shows the result of the proposed model compared with the existing methods. It can be seen that better performance can be achieved than the LSTM models [29].

Table 2. Results on Large-Scale Combined (LSC) dataset.

Method	Cross Sample		Cross Subject	
	Precision	Recall	Precision	Recall
HON4D [50]	84.6%	84.1%	63.1%	59.3%
Dynamic Skeletons [56]	85.9%	85.6%	74.5%	73.7%
AGNN [57]	87.6%	88.1%	84.0%	82.0%
P-LSTM [29]	84.2%	84.9%	76.3%	74.6%
Proposed DA-IndRNN	88.7%	87.3%	80.7%	79.0%

5. Discussion

This paper proposes a new IndRNN-based deep attention model for skeleton-based action recognition. The proposed model employs a deep network to obtain the attention instead of just one or two layers in the conventional attention models, and a deep IndRNN network for classification. Moreover, a new loss function is proposed based on the triplet loss to regulate the attention model among different action categories. The new loss function explicitly enforces the intra-class attention distance to be no larger than the inter-class attention distance. Experiments have demonstrated that the proposed model can attain more robust attention weights and better performance than the existing methods.

The proposed method takes the simple joint coordinates as input, and advanced features such as the geometric features can be considered. In addition, spatial feature extraction techniques such as the graph convolution are worth investigating with the proposed method in the future.

Author Contributions: Conceptualization, Y.G., C.L., S.L., M.Y. and H.Y.; methodology, Y.G., C.L., S.L. and X.C.; software, Y.G. and C.L.; investigation, Y.G., C.L., S.L., X.C., M.Y. and H.Y.; writing, Y.G., C.L. and S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly funded by National Key R&D Program of China (2018YFE0203900), the National Natural Science Foundation of China (No. 61901083, 62001092 and 62101512), Fundamental Research Program of Shanxi Province (20210302124031), the open project program of state key laboratory of virtual reality technology and systems, Beihang University (VRLAB2021A01) and SDU QILU Young Scholars program.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
2. Wang, P.; Li, W.; Gao, Z.; Zhang, Y.; Tang, C.; Ogunbona, P. Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 416–425.
3. Devanne, M.; Wannous, H.; Berretti, S.; Pala, P.; Daoudi, M.; Del Bimbo, A. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Trans. Cybern.* **2015**, *45*, 1340–1352. [CrossRef] [PubMed]
4. Liu, M.; Liu, H.; Chen, C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* **2017**, *68*, 346–362. [CrossRef]
5. Zhang, J.; Han, Y.; Tang, J.; Hu, Q.; Jiang, J. Semi-supervised image-to-video adaptation for video action recognition. *IEEE Trans. Cybern.* **2017**, *47*, 960–973. [CrossRef] [PubMed]
6. Du, W.; Wang, Y.; Qiao, Y. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Volume 2.

7. Girdhar, R.; Ramanan, D. Attentional pooling for action recognition. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 34–45.
8. Wang, P.; Li, W.; Ogunbona, P.; Wan, J.; Escalera, S. RGB-D-based human motion recognition with deep learning: A survey. *Comput. Vis. Image Underst.* **2018**, *171*, 118–139. [CrossRef]
9. Wang, P.; Li, W.; Gao, Z.; Tang, C.; Ogunbona, P.O. Depth Pooling Based Large-Scale 3-D Action Recognition with Convolutional Neural Networks. *IEEE Trans. Multimed.* **2018**, *20*, 1051–1061. [CrossRef]
10. Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; Darrell, T. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
11. Yue-Hei Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
12. Pigou, L.; van den Oord, A.; Dieleman, S.; Van Herreweghe, M.; Dambre, J. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *Int. J. Comput. Vis.* **2018**, *126*, 430–439. [CrossRef]
13. Jain, A.; Zamir, A.R.; Savarese, S.; Saxena, A. Structural-RNN: Deep learning on spatio-temporal graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5308–5317.
14. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building A longer and deeper RNN. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–2 June 2018; pp. 5457–5466.
15. Li, C.; Li, S.; Gao, Y.; Zhang, X.; Li, W. A Two-stream Neural Network for Pose-based Hand Gesture Recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**. [CrossRef]
16. Zhao, B.; Li, S.; Gao, Y.; Li, C.; Li, W. A Framework of Combining Short-Term Spatial/Frequency Feature Extraction and Long-Term IndRNN for Activity Recognition. *Sensors* **2020**, *20*, 6984. [CrossRef]
17. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
18. Sharma, S.; Kiros, R.; Salakhutdinov, R. Action recognition using visual attention. *arXiv* **2015**, arXiv:1511.04119.
19. Song, S.; Lan, C.; Xing, J.; Zeng, W.; Liu, J. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4263–4270.
20. Liu, J.; Wang, G.; Hu, P.; Duan, L.Y.; Kot, A.C. Global context-aware attention lstm networks for 3D action recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1647–1656.
21. Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. In Proceedings of the International Conference on Machine Learning (ICML), Atlanta, GA, USA, 16–21 June 2013.
22. Wang, P.; Li, Z.; Hou, Y.; Li, W. Action recognition based on joint trajectory maps using convolutional neural networks. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; pp. 102–106.
23. Ke, Q.; An, S.; Bennamoun, M.; Sohel, F.; Boussaid, F. SkeletonNet: Mining Deep Part Features for 3-D Action Recognition. *IEEE Signal Process. Lett.* **2017**, *24*, 731–735. [CrossRef]
24. Kim, T.S.; Reiter, A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. *arXiv* **2017**, arXiv:1704.04516.
25. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A New Representation of Skeleton Sequences for 3D Action Recognition. *arXiv* **2017**, arXiv:1703.03492.
26. Cao, C.; Zhang, Y.; Zhang, C.; Lu, H. Body Joint Guided 3-D Deep Convolutional Descriptors for Action Recognition. *IEEE Trans. Cybern.* **2018**, *48*, 1095–1108. [CrossRef] [PubMed]
27. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
28. Veeriah, V.; Zhuang, N.; Qi, G.J. Differential recurrent neural networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4041–4049.
29. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB + D: A large scale dataset for 3D human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
30. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal LSTM with trust gates for 3D human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 816–833.
31. Zhang, S.; Liu, X.; Xiao, J. On geometric features for skeleton-based action recognition using multilayer LSTM networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 148–157.
32. Li, C.; Hou, Y.; Wang, P.; Li, W. Joint Distance Maps Based Action Recognition with Convolutional Neural Networks. *IEEE Signal Process. Lett.* **2017**, *24*, 624–628. [CrossRef]

33. Hou, Y.; Li, Z.; Wang, P.; Li, W. Skeleton Optical Spectra-Based Action Recognition Using Convolutional Neural Networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 807–811. [CrossRef]
34. Liu, J.W.; Akhtar, N.; Mian, A.S. Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition. *arXiv* **2019**, arXiv:1711.05941.
35. Li, B.; He, M.; Dai, Y.; Cheng, X.; Chen, Y. 3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN. *Multimed. Tools Appl.* **2018**, *77*, 22901–22921. [CrossRef]
36. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 2, p. 8.
37. Ba, J.; Mnih, V.; Kavukcuoglu, K. Multiple object recognition with visual attention. *arXiv* **2014**, arXiv:1412.7755.
38. Haque, A.; Alahi, A.; Li, F.-F. Recurrent attention models for depth-based person identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1229–1238.
39. Li, S.; Li, W.; Cook, C.; Gao, Y. Deep Independently Recurrent Neural Network (IndRNN). *arXiv* **2019**, arXiv:1910.06251.
40. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 140–149.
41. Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; Hu, W. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. *arXiv* **2021**, arXiv:2107.12213.
42. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
43. Hoffer, E.; Ailon, N. Deep metric learning using triplet network. In Proceedings of the International Workshop on Similarity-Based Pattern Recognition, Venice, Italy, 28–30 September 2015; pp. 84–92.
44. Zhang, J.; Li, W.; Wang, P.; Ogunbona, P.; Liu, S.; Tang, C. A large scale rgb-d dataset for action recognition. In Proceedings of the International Workshop on Understanding Human Activities through 3D Sensors, Cancun, Mexico, 4 December 2016; pp. 101–114.
45. Huang, Z.; Wan, C.; Probst, T.; Van Gool, L. Deep learning on lie groups for skeleton-based action recognition. *arXiv* **2016**, arXiv:1612.05877.
46. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. *arXiv* **2018**, arXiv:1801.07455.
47. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3D points. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 9–14.
48. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
49. Wang, J.; Liu, Z.; Wu, Y.; Yuan, J. Mining actionlet ensemble for action recognition with depth cameras. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1290–1297.
50. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
51. Yang, X.; Tian, Y. Super normal vector for activity recognition using depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 804–811.
52. Sung, J.; Ponce, C.; Selman, B.; Saxena, A. Human Activity Detection from RGBD Images. In *Plan, Activity, and Intent Recognition*; ACM Digital Library: New York, NY, USA, 2011; Volume 64.
53. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 7–12.
54. Ni, B.; Wang, G.; Moulin, P. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, 6–13 November 2011; pp. 1147–1153.
55. Chen, C.; Jafari, R.; Kehtarnavaz, N. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.

56. Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.
57. Li, C.; Cui, Z.; Zheng, W.; Xu, C.; Ji, R.; Yang, J. Action-Attending Graphic Neural Network. *IEEE Trans. Image Process.* **2018**, *27*, 3657–3670. [CrossRef] [PubMed]