symmetry

Article

PointMapNet: Point Cloud Feature Map Network for 3D Human Action Recognition

Xing Li [1,2], Qian Huang [1,2,*], Yunfei Zhang [1,2], Tianjin Yang [1,2] and Zhijian Wang [1,2]

[1] The Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing 211100, China
[2] School of Computer and Information, Hohai University, Nanjing 211100, China
[*] Correspondence: huangqian@hhu.edu.cn

Abstract: 3D human action recognition is crucial in broad industrial application scenarios such as robotics, video surveillance, autonomous driving, or intellectual education, etc. In this paper, we present a new point cloud sequence network called PointMapNet for 3D human action recognition. In PointMapNet, two point cloud feature maps symmetrical to depth feature maps are proposed to summarize appearance and motion representations from point cloud sequences. Specifically, we first convert the point cloud frames to virtual action frames using static point cloud techniques. The virtual action frame is a 1D vector used to characterize the structural details in the point cloud frame. Then, inspired by feature map-based human action recognition on depth sequences, two point cloud feature maps are symmetrically constructed to recognize human action from the point cloud sequence, i.e., Point Cloud Appearance Map (PCAM) and Point Cloud Motion Map (PCMM). To construct PCAM, an MLP-like network architecture is designed and used to capture the spatio-temporal appearance feature of the human action in a virtual action sequence. To construct PCMM, the MLP-like network architecture is used to capture the motion feature of the human action in a virtual action difference sequence. Finally, the two point cloud feature map descriptors are concatenated and fed to a fully connected classifier for human action recognition. In order to evaluate the performance of the proposed approach, extensive experiments are conducted. The proposed method achieves impressive results on three benchmark datasets, namely NTU RGB+D 60 (89.4% cross-subject and 96.7% cross-view), UTD-MHAD (91.61%), and MSR Action3D (91.91%). The experimental results outperform existing state-of-the-art point cloud sequence classification networks, demonstrating the effectiveness of our method.

Keywords: 3D human action recognition; point cloud sequence; point cloud feature map

Citation: Li, X.; Huang, Q.; Zhang, Y.; Yang, T.; Wang, Z. PointMapNet: Point Cloud Feature Map Network for 3D Human Action Recognition. *Symmetry* **2023**, *15*, 363. https://doi.org/10.3390/sym15020363

Academic Editor: Zhixun Su

Received: 26 December 2022
Revised: 20 January 2023
Accepted: 25 January 2023
Published: 30 January 2023

1. Introduction

The rapid development of low-cost sensors has drawn much attention from computer vision researchers. 3D human action recognition is crucial in application scenarios such as robotics, video surveillance, medical services, autonomous driving, video analysis, or intellectual education, etc. [1]. The depth sequence research [2] is active in the human action recognition area. In the depth sequence, pixel values in depth frames represent the distance information from human bodies to the depth camera. Depth sequences provide the extra body shape, rich 3D structural information, and 3D motion information of the subject in the scene, with insensitivity to lighting conditions, texture, or color changes, and protection of personal privacy. Therefore, we focus on depth sequence-based 3D human action recognition methods.

Early depth sequence-based human action recognition was based on the feature map approach [3], which compressed the entire sequence on one or several feature maps and simultaneously maintained rich spatio-temporal information. After that, some hand-crafted feature extraction methods are applied to the feature maps to extract action features for recognizing actions. With the rapid development of computing capability and the emergence of large-scale datasets [4], deep neural networks have attracted increasing attention and are widely used in action recognition tasks. The combination of depth feature maps and CNNs [5] network architecture is a kind of representative model [6]. Although this network architecture has gained comparable results, it cannot directly learn the action spatio-temporal patterns from depth sequences in the end-to-end learning way. To deal with the relevant limitations, 2D CNNs are extended to 3D CNNs [7] to model spatio-temporal features by using both spatial convolutional and temporal convolution on action sequences. However, 3D CNNs have a large number of model parameters, which directly affect the training time and limit the training performance on a small human action recognition dataset.

To preserve rich spatio-temporal information while reducing computational costs, researchers have recently focused their attention on converting depth sequences into point cloud data. Static point clouds are a collection of points dispersed in three dimensions and can be considered the simplest representation of shapes. Ordered point cloud frames further form a point cloud sequence. Point cloud sequences are a lightweight data type with rich geometry and shape information. To classify point cloud sequences, 3DV-PointNet++ [8] voxelizes point cloud sequences for 3D human action recognition. In 3DV-PointNet++, 3D dynamic voxel (3DV) is proposed as a novel 3D motion representation and input into PointNet++ for extracting the motion features of human action. The appearance features of human action are learned by using the shared PointNet++ to directly model key point cloud frames. However, due to the fact that voxelization is a pre-processing process and is very time-consuming, the 3DV-PointNet++ method is very slow to train and is not an end-to-end deep network model. Therefore, point cloud sequence network methods are necessary, which directly consume the point cloud sequence for 3D human action classification. To directly model point cloud sequences, point spatio-temporal local neighborhoods are constructed, based on which point spatio-temporal operations [9–11] are designed to extract action features. However, point spatio-temporal operations are complex and time consuming and require a large number of parameters to capture the precise appearance and motion features. Moreover, in the point spatio-temporal operations, the spatial information encoding is easily affected by the temporal information encoding, leading to inferior recognition performance.

In this paper, we propose a point cloud feature map-based network called PointMapNet for effective and low-cost 3D human action recognition on point cloud sequences. The core idea of PointMapNet is to utilize two point cloud feature maps to summarize the motion and appearance information in the point cloud sequence for human action recognition. Point cloud feature maps can simplify the point cloud sequence classification task while achieving excellent recognition performance. PointMapNet avoids the computational complexity of voxelization, thus greatly simplifying the point cloud sequence modeling task. In addition, it effectively improves the accuracy of point cloud sequence-based human action recognition by capturing motion and appearance features using two point cloud feature maps, respectively. Specifically, in the data preprocessing stage, we transform the depth sequence data into a point cloud sequence as the input of our PointMapNet. In PointMapNet, we first abstract the point cloud sequence into a virtual action sequence based on the static point cloud technique. Then, we design two parallel point cloud feature map modules with the virtual action sequence and its differential sequence as input to generate PCAM and PCMM, preserving the spatio-temporal appearance and dynamic motion, respectively. Point cloud feature map modules are slightly modified MLP-Mixer structures, in which the MLP-Mixer structure is adopted for inter-frame and intra-frame information communication, and the pooling operations are used to fuse information across frames. Finally, we aggregate the two point cloud feature map representations before performing 3D human action recognition.

Our main contributions are summarized as follows:

- We propose a simple and effective point cloud sequence network, called PointMapNet for 3D human action recognition. PointMapNet is a fully end-to-end optimized network architecture;
- According to our technical contribution, we abstract the point cloud sequences into virtual action sequences, based on which we propose two point cloud feature maps, Point Cloud Appearance Map (PCAM) and Point Cloud Motion Map (PCMM) to obtain the spatio-temporal appearance structure and the motion dynamics for 3D human action recognition;
- Our PointMapNet achieves a cross-view accuracy of 96.7% on the NTU RGB+D 60 dataset, a cross-subject accuracy of 91.61% on the UTD-MHAD dataset, and a cross-subject accuracy of 91.91% on the MSR Action3D dataset, outperforming state-of-the-art methods.

The remainder of the paper is organized as follows. In Section 2, related works are introduced. We present a new point cloud sequence network named PointMapNet in Section 3. In Section 4, experimental results and comparisons are demonstrated and analyzed. In Section 5, the differences from existing methods and limitations of our method are discussed. Finally, conclusions and recommendations are drawn in Section 6.

2. Related Work

2.1. Feature Map-Based Human Action Recognition on Depth Sequence

Depth sequence-based human action recognition approaches usually capture the motion information by computing depth feature maps. Researchers use one or several feature maps instead of an entire depth sequence to study human actions. Depth feature maps are a compact and effective representation used to characterize human actions. Bobick et al. [3] extract masks of human shapes from video sequences, constructing Motion History Images (MHI) and Motion Energy Images (MEI). Wang et al. [12] employ random occupancy patterns to extract semi-local features. Yang et al. [2] propose the Depth Motion Map (DMM) as a representation of human action, where the differences between every two consecutive frames are accumulated over time and then HOGs are calculated from the DMM. Liu et al. [13] present Gabor filters to encode texture data from Adaptive Hierarchical Depth Motion Maps (AH-DMMs) to extract motion and shape cues. Aouaidjia et al. [6] propose the Depth Motion Image (DMI), which assembles depth frames of action to capture the change in the depth sequences of human motion.

2.2. Deep Learning for Static Point Clouds

Our work uses the static point cloud technique to abstract the point cloud sequence into a virtual action sequence. Existing 3D shape classification methods of static point clouds can be classified as multi-view based [14], voxel-based [15], and point-based methods [16–21]. Multi-view based approaches first project a 3D shape into multiple views and extract view-side features. Then, these features are fused for accurate shape classification. Voxel-based methods typically voxelize the point cloud into a 3D mesh and then apply a 3D convolutional neural network (CNN) to the shape classification of the volume representation. Depending on the network structure used for feature learning at each point, the class of point-based methods can be divided into pointwise MLP [16,17], convolution-based [18], graph-based [19,20], and other typical methods [21]. PointNet [16], a pioneering static point cloud work that deals directly with point sets, is a simple, efficient, and powerful feature extractor. PointNet++ [17], a modified method of PointNet, is a hierarchical network that obtains local-global features of different sizes and integrates local neighborhoods with sampling and grouping. In SpiderCNN [18], SpiderConv is presented to define the convolution as the product of a step function and a Taylor expansion defined over *k* nearest neighbors. 3D deep learning work DGCNN [19] proposes to exploit the local geometric structure from a set of 3D points in each local neighborhood. In LDGCNN [20], the transformation network is removed and the layered features of different layers in DGCNN [19] are linked to improve performance and reduce model size. In Ref. [21], the PointSIFT module is integrated into the network to capture the orientation information of the point cloud, achieving strong robustness for shape scaling.

2.3. Deep Learning for Point Cloud Sequences

Early methods extract manual features from point cloud sequences for human action recognition [22,23]. Currently, researchers are focusing on developing deep learning architectures to model point cloud sequences, which yield better recognition performance. Deep learning for point cloud sequence-based 3D human action recognition is a fairly new and challenging task. One solution is to voxelize the point cloud sequence and then transform the voxelized point cloud sequence into existing data to perform the modeling task indirectly [8,24,25]. Fast and Furious (FaF) [24] converts a point cloud frame into a bird's view voxel and then extracts features via 3D convolutions. MinkowskiNet [25] transforms the voxelized point cloud sequence into grid-based videos and then jointly captures the 4D spatio-temporal structure via a 4D convolution kernel. 3DV-PointNet++ [8] converts the voxelized point cloud sequence into static point clouds and then uses a PointNet++ model for human action recognition. Voxel-based point cloud sequence methods have extremely complex data pre-processing operations and cannot model point cloud sequences in an end-to-end manner. Another solution is to develop point cloud sequence networks [9–11], which directly consume point cloud sequences for action recognition. MeteorNet [9] proposes the concept of the spatial-temporal neighborhood by creating direct grouping and chain flow grouping to capture the dynamic points of point cloud sequences. In Ref. [10], a point spatio-temporal (PST) convolution is proposed to achieve informative representations of point cloud sequences. Fan et al. [11] propose a novel Point 4D Transformer (P4Transformer) network including a point 4D convolution to model raw point cloud videos. However, all these point cloud sequence networks have to construct 4D spatio-temporal local neighborhoods to capture the action features. Constructing 4D spatio-temporal local neighborhoods and performing point spatio-temporal operations are computationally complex due to the disordered property of point clouds. In the point spatio-temporal operations, the spatial information encoding is easily affected by the temporal information encoding, leading to inferior recognition performance. In addition, a large number of parameters are necessary for these point cloud sequence networks to obtain fine appearance and motion features. To improve recognition performance and simplify the task of point cloud sequence classification, we propose two point cloud feature maps inspired by depth feature maps to effectively and efficiently characterize the appearance and motion information from point cloud sequences. Based on point cloud feature maps, we propose a lightweight end-to-end point cloud sequence network called PointMapNet, which simplifies the point cloud sequence classification task while achieving impressive recognition performance.

3. Methodology

In this section, we propose a simple and effective point cloud sequence network called PointMapNet for 3D human action recognition. By designing Point Cloud Appearance Map (PCAM) and Point Cloud Motion Map (PCMM) to model the appearance features and motion features of human actions, we explore the spatio-temporal learning capability of PointMapNet on dynamic point cloud sequences. Figure 1 shows the overall flowchart of the proposed framework. Specifically, in the data preprocessing stage, the depth sequence is transformed into a point cloud sequence as the input of our PointMapNet. In PointMapNet, we first abstract the point cloud sequence into a virtual action sequence based on the static point cloud technique. Then, two parallel point cloud feature map modules are designed to generate PCAM and PCMM, preserving the dynamic motion and spatio-temporal appearance, respectively. Finally, we fuse the two point cloud feature map descriptors and send them into a fully-connected classifier for recognizing human actions.

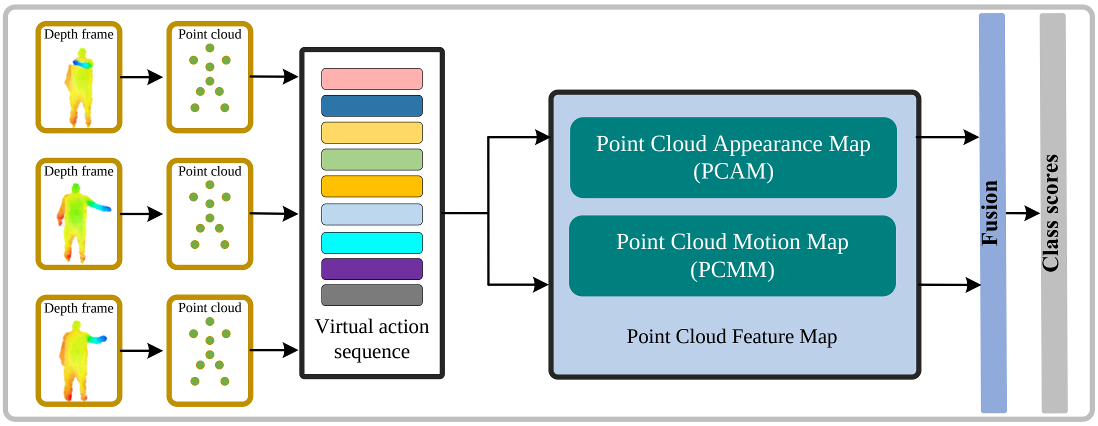

Figure 1. The overall flowchart of the proposed PointMapNet.

3.1. Data Preprocessing

Depth sequences provide the extra body shape, rich 3D structural information, and 3D motion information of the subject in the scene, with insensitivity to lighting conditions, texture or color changes, and protection of personal privacy. Point cloud sequences also have the advantages of depth sequences, while the 3D information is more prominent and the computational cost is lower. Therefore, utilizing point cloud sequences as input for human action recognition is more efficient.

Since the original depth sequence contains a large number of redundant frames, we use a frame selection strategy to select a set of depth frames with the same time interval as input. Specifically, we divide each original depth sequence into T segments of equal length. In the training phase, we randomly select one frame in each segment. In the test phase, we take one frame at a fixed time position in each segment. Then, the depth sequence of T frames is converted to an original point cloud sequence by converting the pixel position, and the pixel value in the selected point cloud frames to the 3D coordinates (X, Y, Z). We randomly select N points in each original point cloud frame to form the final point cloud sequence $\boldsymbol{S} = \{S_t\}_{t=1}^{T}$. $S_t = \{x_t^1, x_t^2 \ldots, x_t^N\}$ represents the unordered point set of the t-th point cloud frame.

3.2. Virtual Action Sequence

Each frame of the point cloud sequence describes the spatial distribution of the static appearance of the human body. In this paper, we use the virtual action frames as a static summary of the appearance information in the point cloud frames. The core idea of PointMapNet is point cloud feature maps, which represent action spatio-temporal features by collecting the appearance and motion information from virtual action sequences. To this end, the static point cloud technique is used to abstract point cloud sequences into virtual action sequences.

To generate the virtual action sequence, we first perform twice set abstraction operations [17] to downsample the point clouds in each frame. Then, the PointNet layer [16] is used to merge the information of all the downsampled points to form the virtual action frame. Finally, T virtual action frames constitute the virtual action sequence $\boldsymbol{V} \in \mathbb{R}^{T \times C}$.

3.3. Point Cloud Appearance Map (PCAM) Module

Review on Depth Motion Image (DMI): As shown in Figure 2a, DMI describes the spatio-temporal appearance of actions by finding the most prominent appearance features from the same spatial location in all depth frames, producing a unified representation that defines each action with its own specific spatio-temporal appearance, and providing unique features for each action.

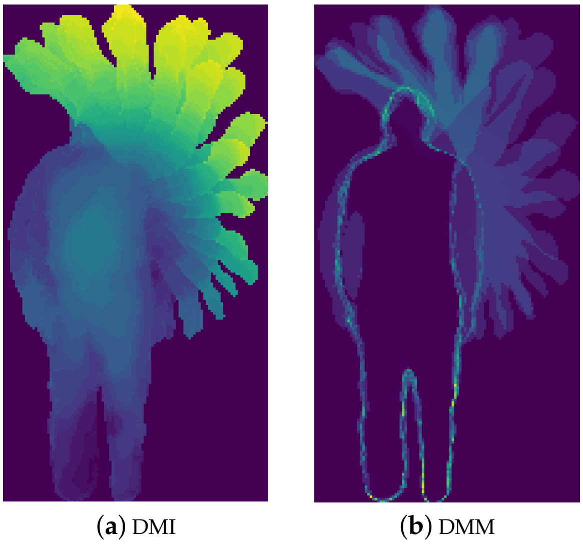
(**a**) DMI (**b**) DMM

Figure 2. Depth feature maps.

The following equation illustrates the calculation of DMI:

$$\mathrm{DMI}(i,j) = \underset{t=1,\ldots,T}{\mathrm{MAX}}(I(i,j,t)) \tag{1}$$

where $I(i,j,t)$ is the pixel value at tth depth frame on pixel position (i,j). The range of t is from frame 1 to T, T denotes the total number of frames. The pixel value of the DMI is the maximum value of the same pixel position, which represents the most prominent action appearance feature in the depth sequence.

Point Cloud Appearance Map: Inspired by DMI, PCAM describes spatial-temporal appearance by finding the most prominent features at the same spatial location of all static appearances for 3D human action recognition. A PCAM module based on the slightly modified MLP-Mixer [26] is designed to aggregate the appearance information in virtual action frames.

The PCAM module is a slightly modified MLP-Mixer structure shown in Figure 3 with the virtual action sequence $\boldsymbol{V} \in \mathbb{R}^{T\times C}$ as input. It contains an MLP-Mixer structure of two types of MLP blocks and one pooling operation. In the PCAM module, the MLP-Mixer structure is used to enhance the human appearance information in virtual action frames, and the max-pooling operation is adopted to select the most remarkable appearance features along the time dimension to form the PCAM descriptor.

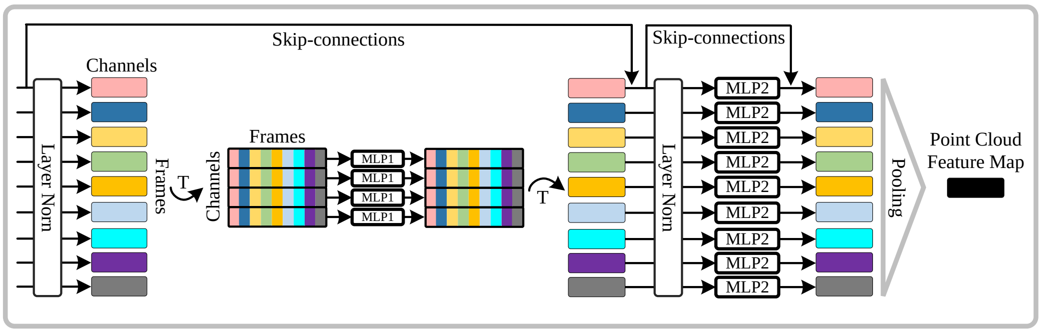

Figure 3. The network structure of the slightly modified MLP-Mixer.

The first MLP block in the PCAM module is the inter-frame MLP block, which allows information communication between different virtual action frames. The inter-frame MLP block acts on columns of $\boldsymbol{V}$ (i.e., the transposition of input $\boldsymbol{V}^{\top}$), maps $\mathbb{R}^{T} \rightarrow \mathbb{R}^{T}$, and is shared across all columns. Each inter-frame MLP block consists of two fully-connected layers, a non-linearity, and a skip-connection. The inter-frame MLP can be formalized as follows:

$$\boldsymbol{Y}_{*,d} = \boldsymbol{V}_{*,d} + \boldsymbol{W}_2\sigma(\boldsymbol{W}_1 \,\mathrm{LayerNorm}(\boldsymbol{V})_{*,d}), \quad d = 1 \dots D \tag{2}$$

where σ is an element-wise nonlinearity (GELU). $\boldsymbol{W}_1, \boldsymbol{W}_2$ are fully connected layers. LayerNorm imposes additional constraints on the distribution of the data, thus enhancing the generalization capability of the model. $\boldsymbol{Y}$ is the intermediate feature matrix of the virtual action sequence. The skip-connection solves the problem of gradient disappearance during the training process.

Then, the second MLP block, i.e., the intra-frame MLP block, acts on each intermediate feature to refine the intra-frame appearance information. The intra-frame MLP block acts on rows of $\boldsymbol{Y}$, maps $\mathbb{R}^T \to \mathbb{R}^T$, and is shared across all virtual action frames. The intra-frame MLP block can be written as follows:

$$\boldsymbol{M}_{t,*} = \boldsymbol{Y}_{t,*} + \boldsymbol{W}_4\sigma(\boldsymbol{W}_3 \,\mathrm{LayerNorm}(\boldsymbol{Y})_{t,*}), \quad t = 1 \dots T \tag{3}$$

where $\boldsymbol{W}_3, \boldsymbol{W}_4$ are fully connected layers in the intra-frame MLP block. $\boldsymbol{M}$ denotes the updated feature matrix of the virtual action sequence. Finally, the PCAM module uses a global max-pooling layer to select the most remarkable appearance features along the time dimension to form the PCAM descriptor. To simplify the formula of the PCAM module, we use the function MLP-Mixer($\cdot$) instead of the two MLP blocks. The calculation procedure of $\boldsymbol{M}$ can be rewritten as:

$$\boldsymbol{M}_t = \text{MLP-Mixer}(\boldsymbol{V}_t) \tag{4}$$

Based on Equation (4), the PCAM module can be formalized as follows:

$$\begin{aligned} PCAM_d &= \underset{t=1,\dots,T}{\mathrm{MAX}}\{\boldsymbol{M}_{t,d}\} \\ &= \underset{t=1,\dots,T}{\mathrm{MAX}}\{\text{MLP-Mixer}(\boldsymbol{V}_t)_d\} \end{aligned} \tag{5}$$

where d is the channel position. $PCAM$ is the PCAM descriptor.

3.4. Point Cloud Motion Map (PCMM) Module

Review on Depth Motion Map (DMM): As shown in Figure 2b, the DMM describes the accumulated motion distribution and intensity of the action. To calculate the DMM, the motion energy is first acquired by computing the difference between two consecutive depth video frames. Then, the motion energy is stacked over the entire video sequence. The depth feature map DMM for depth video sequences can be represented as:

$$DMM(i,j) = \sum_{t=1}^{T-1}(|I(i,j,t+1) - I(i,j,t)|) \tag{6}$$

The DMM reflects the accumulation of motion energy in human actions. It provides a strong cue to the category of motion in progress.

Point Cloud Motion Map: The PCMM inspired by DMM captures dynamic movements by accumulating all motion energy over the whole time for 3D human action recognition. In order to obtain the motion characteristics, the PCMM sequentially calculates the motion energy between two virtual action frames. The motion energy sequence $\boldsymbol{E} = \{E_t\}_{t=1}^{T-1}$ is different from the virtual action sequence $\boldsymbol{V}$, which computes the difference between two consecutive virtual action frames:

$$E_t = \{|V_{t+1} - V_t|\} \quad t = 1, 2, \dots, T-1 \tag{7}$$

The PCMM module still uses the slightly modified MLP-Mixer to enhance and aggregate motion feature information. Unlike the PCAM module, a sum-pooling operation is used instead of the max-pooling operation in the PCMM module:

$$\begin{aligned} PCAM_d &= \textstyle\sum_{t=1}^{T-1}(\boldsymbol{M}_{t,d}) \\ &= \textstyle\sum_{t=1}^{T-1}(\text{MLP-Mixer}(E_t)_d) \\ &= \textstyle\sum_{t=1}^{T-1}(\text{MLP-Mixer}(|V_{t+1}-V_t|)_d) \end{aligned} \tag{8}$$

where $\boldsymbol{M}$ denotes the updated feature matrix in the PCMM module. $PCMM$ is the PCMM descriptor.

3.5. Two Maps Fusion

To generate the final feature of each action sequence, we simply concatenate the descriptors of two point cloud feature maps, including the PCAM descriptor and the PCMM descriptor. Finally, we use a set of fully connected layers as a classifier for 3D action recognition.

4. Experiments

In this section, we introduce the datasets and experimental implementation details. We choose the three most widely used 3D human action recognition datasets to evaluate the performance of our proposed method, NTU RGB-D 60 [4], UTD-MHAD [27], and MSR Action3D [28]. Then, we compare PointMapNet with existing state-of-the-art methods and perform ablation studies to further validate the contributions of different components in PointMapNet. Finally, we compare the running time of our PointMapNet and 3DV-PointNet++ on CPU and GPU.

4.1. Datasets

We conduct extensive experiments on a large-scale public dataset (i.e., NTU RGB+D 60 [4]) and two small-scale public datasets (i.e., UTD Multimodal Human Action Dataset [27] (UTD-MHAD), and MSR Action3D [28]).

The NTU RGB+D 60 dataset contains RGB sequences, infrared sequences, depth sequences, and skeleton point sequences, including over 56,000 video sequences. This dataset includes 4 million frames performed by 40 different subjects from 60 different action classes. Two protocols, namely the cross-view (CV) evaluation protocol and the cross-subject (CS) evaluation protocol, are defined for performance evaluation.

The UTD Multimodal Human Action Dataset (UTD-MHAD) consists of 861 depth video sequence samples of 27 actions performed by 8 subjects. Each subject performs each action 4 times. We use the same experimental setting as in [27], half of the subjects are used for training, and the other half for testing.

The MSR Action3D dataset is one of the most classical action recognition datasets, containing 567 sequences from 20 different human actions performed by 10 subjects. The same cross-subject test setting as [28] is adopted. in which odd subjects (1, 3, 5, 7, and 9) are used for training and even subjects (2, 4, 6, 8, and 10) are used for testing. In Figure 4, the point cloud sequence of horizontal arm wave in the MSR Action3D dataset is shown as an example of action samples.

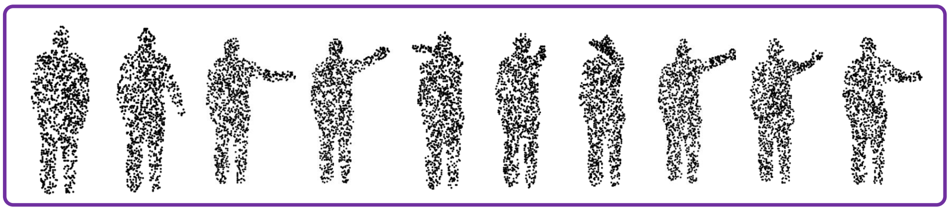

Figure 4. The point cloud sequence of horizontal arm wave in the MSR Action3D dataset.

4.2. Implementation Details

The implementation details about our network architecture are provided as follows. First, 512 points are randomly selected from each original point cloud frame to generate the point cloud sequence. Then, in the virtual action sequence generation stage, we perform two set abstraction operations on each point cloud frame to obtain the downsampled point set. In the first set abstract operation, 128 centroid points are selected to determine point groups, and the output channels of MLPs are set to 64, 64, and 128, respectively. The group radius is set to 0.06. The point number k_1 in each point group is set to 48. In the second set abstract operation, 32 centroid points are selected to determine point groups, and the output channels of MLPs are set to 128, 128, and 256, respectively. The group radius is set to 0.1. The point number k_2 in each point group is set to 16. A PointNet layer is used to integrate all downsampled points, in which the output channels of MLPs are set to 256, 512, and 1024, respectively. In the point cloud feature map generation stage, two groups of MLPs are used in the inter-frame MLP block and intra-frame MLP block. The output channels of the inter-frame MLP block are set to 1024 and 1024, respectively. The output channels of the intra-frame MLP block are set to 1024 and 1024, respectively. The output channels of the final set of MLPs used as the classifier are set as 256 and the number of action categories.

4.3. Comparison with the State-of-the-Art Methods

4.3.1. NTU RGB+D 60 Dataset

To validate the performance of the proposed network, we compared PointMapNet with the state-of-the-art method on the NTU RGB+D 60 dataset. In Table 1, we compare PointMapNet with the depth sequence methods and the skeletal sequence methods. The depth sequence methods include HON4D [29], Multi-view Dynamic Images [30], HDDPDI [31], Dynamic images (HRP) [32], 3DFCNN [33], Stateless ConvLSTM [34], and 3DV-PointNet++ [8]. The skeletal sequence methods include VA-CNN [35], SGN [36], DGNN [37], DC-GCN+ADG [38], and ST-GCNs [39]. The recognition results of the proposed PointMapNet are tabulated in Table 1, which shows that our approach yields a perfect recognition accuracy of 96.7% and 89.4% under the cross-view setting and cross-subject setting, respectively. Impressively, PointMapNet results in the highest recognition accuracy among all depth sequence methods under the cross-view setting. Compared to the methods with depth sequences as input, PointMapNet transforms the depth sequence into a point cloud sequence to enhance the 3D shape structure of human actions. Compared to the method with point cloud sequences as input, PointMapNet directly consumes point cloud sequences instead of voxelized point cloud sequences for human action recognition, which avoids the expensive computational cost of voxelization operations. PointMapNet proposes two point cloud feature maps to effectively extract the spatio-temporal appearance and aggregate the motion energy of human actions. Furthermore, the performance of our PointMapNet is the second-best method under the cross-subject setting, slightly lower than DGNN by 0.5%. Compared to the skeletal sequence approaches, PointMapNet does not require extra pose estimation algorithms and the point cloud sequence has more details of the body structure than the skeletal sequence.

Table 1. Action recognition accuracy (%) on the NTU RGB+D 60 dataset.

Method	Input	CS	CV
HON4D [29]	Depth	30.6	7.3
Multi-view Dynamic Images [30]	Depth	84.6	87.3
HDDPDI [31]	Depth	82.4	87.6
Dynamic images (HRP) [32]	Depth	87.1	84.2
3DFCNN [33]	Depth	78.1	80.3
Stateless ConvLSTM [34]	Depth	75.3	75.5
3DV-PointNet++ [8]	Depth/Point	88.8	96.3
VA-CNN [35]	Skeleton	88.7	94.3
SGN [36]	Skeleton	89.0	94.5
DGNN [37]	Skeleton	89.9	96.1
DC-GCN+ADG [38]	Skeleton	88.2	95.2
ST-GCNs [39]	Skeleton	81.5	88.3
PointMapNet (ours)	Depth/Point	89.4	96.7

4.3.2. UTD-MHAD Dataset

To comprehensively measure the performance of our network, we perform comparative experiments on the small-scale UTD-MHAD dataset. In Table 2, we compare PointMapNet with the depth sequence method. The depth sequence methods include Baseline [27], DMM-HOG [2], 3DHOT-MBC [40], DMI [6], HP-DMM-CNN [41], HP-DMM-HOG [42], DSIEM [43], and DRDIS [44]. As shown in Table 2, PointMapNet proposed in this paper achieves a recognition accuracy of 91.61%, which is superior to other state-of-the-art methods. The key success of our PointMapNet lies in the extra spatial shape information of human movements provided by the point cloud sequences and the powerful action characterization capability of the point cloud feature maps. In contrast to the manual feature maps in depth sequence approaches, the point cloud feature maps are automatically learned through network optimization, enhancing the ability to characterize action sequences.

Table 2. Action recognition accuracy (%) on the UTD-MHAD dataset.

Method	Input	Accuracy
Baseline [27]	Depth	66.1
DMM-HOG [2]	Depth	81.5
3DHOT-MBC [40]	Depth	84.4
DMI [6]	Depth	82.79
HP-DMM-CNN [41]	Depth	82.75
HP-DMM-HOG [42]	Depth	73.72
DSIEM [43]	Depth	88.37
DRDIS [44]	Depth	87.88
PointMapNet (ours)	Depth/Point	91.61

4.3.3. MSR Action3D Dataset

Another small-scale MSR Action3D dataset is adopted to evaluate the performance of our network. In Table 3, we compare PointMapNet with the depth sequence method. The depth sequence methods include Baseline [28], HON4D [29], STOP [12], DSTIP [45], PointNet++ [17], MeteorNet [9], PSTNet [10], and P4Transformer [11]. As shown in Table 3, PointMapNet has the highest recognition accuracy of 91.91% when 24 frames are selected from the depth map sequence, which proves the superiority of the proposed PointMapNet. The two-stream architecture in PointMapNet decomposes the appearance structure encoding and motion energy encoding in the point cloud sequence, which avoids the mutual interference between temporal and spatial information. Similar to our method, MeteorNet, PSTNet, and P4Transformer are also point cloud sequence models that classify point cloud sequences in an end-to-end manner. These approaches construct 4D spatio-temporal local neighborhoods and then perform point spatio-temporal operations to capture the point dynamics in point cloud sequences. However, constructing spatio-temporal local neighborhoods inevitably confuses the appearance and motion information and is computationally complex. The experimental results on two small-scale databases illustrate that our method has the ability to achieve outstanding performance even with a small number of samples for training.

Table 3. Action recognition accuracy (%) on the MSR Action3D dataset.

Method	Input	# of Frames	Accuracy
Baseline [28]	Depth	full	74.70
HON4D [29]	Depth	full	88.89
STOP [12]	Depth	full	84.8
DSTIP [45]	Depth	full	89.3
PointNet++ [17]	Depth	1	61.61
MeteorNet [9]	Depth/Point	4	78.11
		8	81.14
		12	86.53
		16	88.21
		24	88.50
PSTNet [10]	Depth/Point	4	81.14
		8	83.50
		12	87.88
		16	89.90
		24	91.20
P4Transformer [11]	Depth/Point	4	80.13
		8	83.17
		12	87.54
		16	89.56
		24	90.94
PointMapNet (ours)	Depth/Point	4	79.04
		8	84.93
		12	87.13
		16	89.71
		20	91.18
		24	91.91

4.4. Ablation Study

In this section, detailed ablation experiments are performed to validate the contributions of different components of PointMapNet. We use the NTU RGB+D 60 dataset as an example to perform comprehensive ablation studies with the cross-view evaluation protocol.

4.4.1. Contributions of Point Cloud Feature Maps in Our Framework

We first investigate the contribution of two different point cloud feature maps in PointMapNet, and the results are reported in Table 4. Inspired by DMI, we propose PCAM to describe the spatio-temporal appearance information in point cloud sequences. Inspired by DMM, we present PCMM to characterize the motion energy information of human actions. The PCAM-only PointMapNet is denoted as PointMapNet (w/o PCMM), i.e., only the PCAM module network is used and the PCMM module network is not included. By contrast, the PointMapNet using only PCMM is denoted as PointMapNet (w/o PCAM), i.e., it uses only the PCMM module network and does not include the PCAM module network. As shown in Table 4, PointMapNet (w/o PCMM) achieves an accuracy of 95.5%, which indicates that the prominent spatio-temporal appearance information enhanced by the MLP-Mixer structure and selected by the max-pooling operation in PCAM module is useful for 3D action recognition. PointMapNet (w/o PCAM) also obtains an accuracy of 86.6%. The experimental results show that our proposed PCAM and PCMM can still achieve competitive performance when used independently. Moreover, the accuracy of PointMapNet (w/o PCMM) is 8.9% higher than that of PointMapNet (w/o PCAM), which indicates that spatio-temporal appearance features are more important for action recognition than motion energy features.

Table 4. Cross-view recognition accuracy (%) when using single point cloud feature map on the NTU RGB+D 60 dataset.

Method	Accuracy
PointMapNet (w/o PCMM)	95.5
PointMapNet (w/o PCAM)	86.6
PointMapNet	96.7

4.4.2. Effectiveness of the MLP-Mixer Network Structure

PointMapNet contains a PCAM module and a PCMM module, both of which use a slightly modified MLP-Mixer structure for inter-frame and intra-frame information communication. In the PCAM module, the MLP-Mixer structure is used to enhance the human appearance information in virtual action frames and to form the PCAM descriptor by selecting the most remarkable appearance features along the time dimension using a max-pooling operation. In the PCMM module, the MLP-Mixer structure is employed to reinforce the human motion information in motion energy frames and to form the PCMM descriptor by aggregating the motion features in all time dimensions using a sum-pooling operation. To verify the performance of the slightly modified MLP-Mixer structure, we propose a baseline approach that computes the PCAM descriptor and the PCMM descriptor directly using the max-pooling and sum-pooling operation instead of the modified MLP-Mixer structure. The PointMapNet using only the max-pooling operation is denoted as PointMapNet (max-pooling), and its input is the virtual action sequence. The PointMapNet using only the sum-pooling operation is denoted as PointMapNet (sum-pooling), and its input is the motion energy sequence. The PointMapNet using both max-sum and pooling operations is denoted as PointMapNet (pooling), which is a two-stream structure such as PointMapNet. The results are listed in Table 5. From the table, it can be seen that the recognition accuracy of PointMapNet (max-pooling) is 90.4%, PointMapNet (sum-pooling) is 82.6%, and PointMapNet (pooling) is 92.8%. The experimental results show that the slightly modified MLP-Mixer structure can effectively enhance inter-frame and intra-frame information communication, thereby enhancing the appearance and motion information.

Table 5. Comparison of cross-view recognition accuracy (%) without the MLP-Mixer structure on the NTU RGB+D 60 dataset.

Method	Accuracy
PointMapNet (max-pooling)	90.4
PointMapNet (sum-pooling)	82.6
PointMapNet (pooling)	92.8

4.4.3. Effectiveness of Different Point Cloud Frame Numbers

We choose the different number of point cloud frames (i.e., T = 4, 8, 12, 16, 20, 24, 28) to verify the effect of the number of point cloud frames on the accuracy of 3D human action recognition. As shown in Figure 5, the recognition accuracy of PointMapNet gradually improves as the number of point cloud frames increases. After the number of point cloud frames reaches 24, the recognition accuracy of PointMapNet tends to be stable. Considering that the running time of PointMapNet also increases with the number of point cloud frames, we choose a point cloud sequence of 24 frames as the input of PointMapNet, and thus achieve the best balance of accuracy and running time.

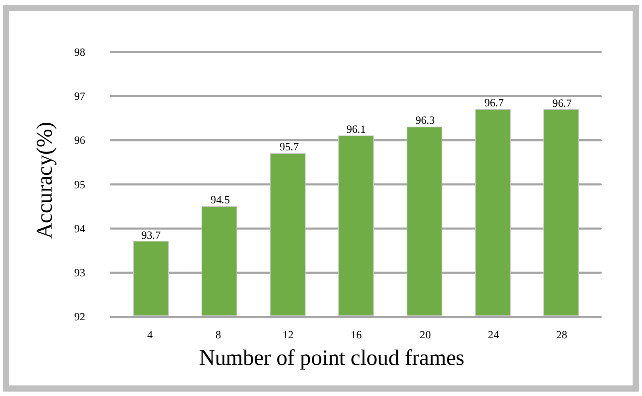

Figure 5. Recognition accuracy (%) of our PointMapNet when using different numbers of frames on the NTU RGB+D 60 dataset.

4.5. Visualization on Different Types of Depth Cloud Feature Maps

To show the discriminative capability of different types of point cloud feature map features, the t-SNE method [46] is used for feature visualization in Figure 6. We compare the PCAM descriptor feature, the PCMM descriptor feature, and their joint features. We use the MSR Action3D dataset as an example to visualize the different types of features from the first eight samples. As shown in Figure 6, we observe that the joint features of the PCAM descriptor and the PCMM descriptor have smaller intra-class distances and larger inter-class distances, which indicates that it is beneficial to integrate the two point cloud feature maps.

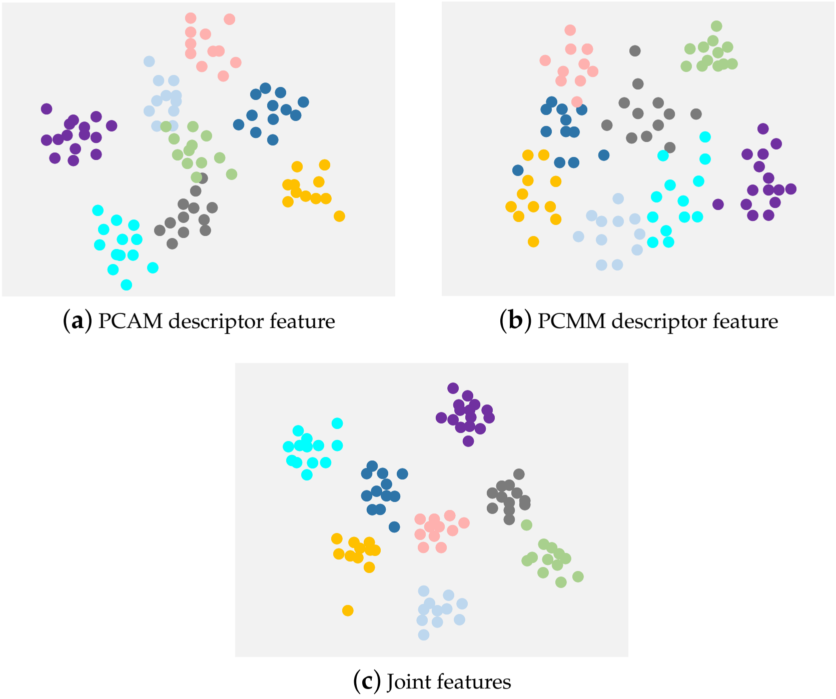

Figure 6. The t-SNE visualization results based on different types of point cloud feature map features from the samples in the first eight classes on the MSR Action3D dataset.

4.6. Running Time and Memory Usage Analysis

To verify the computational efficiency of PointMotionNet, we compare the running time and memory usage of our method with that of 3DV-PointNet++. The model running time includes CPU-based data generation time and GPU-based network forward inference time. The comparison experiments are performed using an Intel(R) Xeon(R) W-3175X CPU and an Nvidia RTX 3090 GPU on the NTU RGB+D 60. Our PointMapNet is much faster than 3DV-PointNet++ in terms of data generation speed and network forward inference speed in Table 6. In addition, our method has slightly more parameters than 3DV-PointNet++. This is because 3DV-PointNet++ converts the point cloud sequence to a static point cloud. Encoding a static point cloud is simpler than encoding a point cloud sequence. However, this advantage is predicated on the fact that 3DV-PointNet++ sacrifices the ability to model point cloud sequences in an end-to-end manner and requires a complicated voxelization operation.

Table 6. Running time (milliseconds per point cloud sequence) and memory usage (M) of 3DV-PointNet++ and our PointMapNet.

Method	CPU Time	GPU Time	Total Time	Parameters
3DV-PointNet++	4842	54	4896	1.24
PointMapNet	2426	15	2441	2.65

5. Discussion

Recent studies are usually based on depth sequences, skeleton sequences, and point cloud sequences for 3D human action recognition. Compared with depth sequences, point cloud sequences are simpler to encode. Compared with skeleton sequences, point cloud sequences have richer structural details and do not require extra pose estimation algorithms to generate the input data. This paper focuses on 3D human action recognition based on point cloud sequences and proposes a new point cloud sequence network called PointMapNet. Extensive experimental results show that PointMapNet achieves state-of-the-art results with excellent operational efficiency on three widely used human action recognition datasets. Deep learning on point cloud sequences can be categorized as voxel-based approaches and point cloud sequence network approaches. Voxel-based approaches [8,25] transform voxelized point cloud sequences into existing data type and model them using the corresponding techniques. Voxel-based methods avoid extracting features directly from complex point cloud sequences. However, the voxelization operation is computationally intensive and loses a lot of spatio-temporal detail information. In addition, voxel-based methods cannot model point cloud sequences in an end-to-end manner. Point cloud sequence network approaches [9–11] consume point cloud sequences directly and encode point cloud sequences in an end-to-end manner. Compared with existing point cloud sequence network methods, our approach proposes two point cloud feature maps to characterize the action features, avoiding the computational complexity of point spatio-temporal operations and thus simplifying the point cloud sequence classification task. In addition, the two-stream network architecture in PointMapNet decomposes temporal information encoding and spatial information encoding, which prevents the mutual interference between appearance and motion features.

In general, the 3D human action recognition based on PointMapNet is very effective. However, there are also some limitations in PointMapNet. PointMapNet cannot effectively capture the long-term relationships in the point cloud sequences. As shown in Table 4, the ability of the PCMM module in PointMapNet to capture motion information is insufficient. In addition, PointMapNet is only applicable to the classification task of point cloud sequences but not to the semantic segmentation task, which limits the generality of PointMapNet.

6. Conclusions and Recommendation

In this paper, we propose a novel point cloud sequence network called PointMapNet for 3D human action recognition. In the point cloud sequence modeling process, it is extremely challenging to avoid computationally complex point spatio-temporal operations and the mutual interference between appearance and motion features. In PointMapNet, we creatively design two kinds of point cloud feature maps to capture the appearance and motion information in point cloud sequences, i.e., Point Cloud Appearance Map (PCAM) and Point Cloud Motion Map (PCMM). Comprehensive experiments performed on three widely used public datasets demonstrate the effectiveness of our approach. We obtain the conclusion that PointMapNet based on point cloud feature maps effectively avoids computationally expensive point spatio-temporal operations, thereby simplifying the task of point cloud sequence classification. Moreover, PointMapNet decomposes the appearance and motion information encoding, preventing the mutual interference between them, thus improving the action recognition accuracy.

For future work, we intend to investigate more efficient static point cloud encoding techniques to obtain more fine-grained virtual action sequences. In addition, how to improve the ability to capture motion features is also planned to be studied.

Author Contributions: Conceptualization, X.L. and Q.H.; methodology, X.L.; software, X.L.; validation, X.L., T.Y. and Y.Z.; formal analysis, X.L.; investigation, Y.Z.; resources, Q.H.; data curation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L.; visualization, T.Y.; supervision, Z.W.; project administration, Q.H.; funding acquisition, Q.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018YFC0407905, the 14th Five-Year Plan for Educational Science of Jiangsu Province under grant no. D/2021/01/39, the Jiangsu Higher Education Reform Research Project under grant no. 2021JSJG143, and the 2022 Undergraduate Practice Teaching Reform Research Project of Hohai University.

Data Availability Statement: The data sets used in this paper are public, free, and available at. NTU RGB+D 60, NTU RGB+D 120: https://rose1.ntu.edu.sg/dataset/actionRecognition/; (accessed on 21 August 2021).

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

References

1. Yang, W.; Zhang, J.; Cai, J.; Xu, Z. Relation Selective Graph Convolutional Network for Skeleton-Based Action Recognition. *Symmetry* **2021**, *13*, 2275. [CrossRef]
2. Yang, X.; Zhang, C.; Tian, Y. *Recognizing Actions Using Depth Motion Maps-Based Histograms of Oriented Gradients*; Association for Computing Machinery: New York, NY, USA, 2012. [CrossRef]
3. Bobick, A.; Davis, J. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 257–267. [CrossRef]
4. 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.
5. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
6. Kamel, A.; Sheng, B.; Yang, P.; Li, P.; Shen, R.; Feng, D.D. Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures. *IEEE Trans. Syst. Man Cybern. Syst.* **2019**, *49*, 1806–1819. [CrossRef]
7. Li, X.; Shuai, B.; Tighe, J. Directional temporal modeling for action recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 275–291.
8. Wang, Y.; Xiao, Y.; Xiong, F.; Jiang, W.; Cao, Z.; Zhou, J.T.; Yuan, J. 3dv: 3d dynamic voxel for action recognition in depth video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 511–520.
9. Liu, X.; Yan, M.; Bohg, J. Meteornet: Deep learning on dynamic 3d point cloud sequences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9246–9255.

10. Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; Kankanhalli, M. PSTNet: Point spatio-temporal convolution on point cloud sequences. *arXiv* **2022**, arXiv:2205.13713.
11. Fan, H.; Yang, Y.; Kankanhalli, M. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 14204–14213.
12. Wang, J.; Liu, Z.; Chorowski, J.; Chen, Z.; Wu, Y. Robust 3d action recognition with random occupancy patterns. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 872–885.
13. Liu, H.; He, Q.; Liu, M. Human action recognition using adaptive hierarchical depth motion maps and gabor filter. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 1432–1436.
14. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-View Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
15. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
16. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
17. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
18. Xu, Y.; Fan, T.; Xu, M.; Zeng, L.; Qiao, Y. SpiderCNN: Deep Learning on Point Sets with Parameterized Convolutional Filters. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
19. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* **2019**, *38*, 146. [CrossRef]
20. Zhang, K.; Hao, M.; Wang, J.; de Silva, C.W.; Fu, C. Linked Dynamic Graph CNN: Learning on Point Cloud via Linking Hierarchical Features. *arXiv* **2019**, arXiv:1904.10014. [CrossRef]
21. Jiang, M.; Wu, Y.; Zhao, T.; Zhao, Z.; Lu, C. PointSIFT: A SIFT-like Network Module for 3D Point Cloud Semantic Segmentation. *arXiv* **2018**, arXiv:1807.00652. [CrossRef]
22. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as Space-Time Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253. [CrossRef] [PubMed]
23. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013.
24. Luo, W.; Yang, B.; Urtasun, R. Fast and Furious: Real Time End-to-End 3D Detection, Tracking and Motion Forecasting With a Single Convolutional Net. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
25. Choy, C.; Gwak, J.; Savarese, S. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
26. Tolstikhin, I.O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. Mlp-mixer: An all-mlp architecture for vision. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24261–24272.
27. 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.
28. 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.
29. 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.
30. Xiao, Y.; Chen, J.; Wang, Y.; Cao, Z.; Tianyi Zhou, J.; Bai, X. Action recognition for depth video using multi-view dynamic images. *Inf. Sci.* **2019**, *480*, 287–304. [CrossRef]
31. Wu, H.; Ma, X.; Li, Y. Hierarchical dynamic depth projected difference images–based action recognition in videos with convolutional neural networks. *Int. J. Adv. Robot. Syst.* **2019**, *16*, 1729881418825093. [CrossRef]
32. 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]
33. Sanchez-Caballero, A.; de López-Diz, S.; Fuentes-Jimenez, D.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Casillas-Perez, D.; Sarker, M.I. 3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information. *Multimed. Tools Appl.* **2022**, *81*, 24119–24143. [CrossRef]
34. Sanchez-Caballero, A.; Fuentes-Jimenez, D.; Losada-Gutiérrez, C. Exploiting the convlstm: Human action recognition using raw depth video-based recurrent neural networks. *arXiv* **2020**, arXiv:2006.07744.
35. Zhang, P.; Lan, C.; Xing, J.; Zeng, W.; Xue, J.; Zheng, N. View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1963–1978. [CrossRef] [PubMed]

36. Zhang, P.; Lan, C.; Zeng, W.; Xing, J.; Xue, J.; Zheng, N. Semantics-guided neural networks for efficient skeleton-based human action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 1112–1121.
37. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7912–7921.
38. Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; Lu, H. Decoupling gcn with dropgraph module for skeleton-based action recognition. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 536–553.
39. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
40. Zhang, B.; Yang, Y.; Chen, C.; Yang, L.; Han, J.; Shao, L. Action Recognition Using 3D Histograms of Texture and A Multi-Class Boosting Classifier. *IEEE Trans. Image Process.* **2017**, *26*, 4648–4660. [CrossRef] [PubMed]
41. Elmadany, N.E.D.; He, Y.; Guan, L. Information Fusion for Human Action Recognition via Biset/Multiset Globality Locality Preserving Canonical Correlation Analysis. *IEEE Trans. Image Process.* **2018**, *27*, 5275–5287. [CrossRef] [PubMed]
42. Rahmani, H.; Mahmood, A.; Du Huynh, Q.; Mian, A. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 742–757.
43. Yang, T.; Hou, Z.; Liang, J.; Gu, Y.; Chao, X. Depth Sequential Information Entropy Maps and Multi-Label Subspace Learning for Human Action Recognition. *IEEE Access* **2020**, *8*, 135118–135130. [CrossRef]
44. Wu, H.; Ma, X.; Li, Y. Spatiotemporal Multimodal Learning With 3D CNNs for Video Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 1250–1261. [CrossRef]
45. Xia, L.; Aggarwal, J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2834–2841.
46. Xie, S.; Sun, C.; Huang, J.; Tu, Z.; Murphy, K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 305–321.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.