applied sciences

MDPI

Article

A Novel Symmetric Fine-Coarse Neural Network for 3D Human Action Recognition Based on Point Cloud Sequences

Chang Li [1], Qian Huang [1,*], Yingchi Mao [1], Weiwen Qian [1] and Xing Li [2]

[1] College of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China; lichang@hhu.edu.cn (C.L.); yingchimao@hhu.edu.cn (Y.M.); qianweiwen@hhu.edu.cn (W.Q.)
[2] College of Information Science and Technology and College of Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China; lixing@njfu.edu.cn
* Correspondence: huangqian@hhu.edu.cn

Abstract: Human action recognition has facilitated the development of artificial intelligence devices focusing on human activities and services. This technology has progressed by introducing 3D point clouds derived from depth cameras or radars. However, human behavior is intricate, and the involved point clouds are vast, disordered, and complicated, posing challenges to 3D action recognition. To solve these problems, we propose a Symmetric Fine-coarse Neural Network (SFCNet) that simultaneously analyzes human actions' appearance and details. Firstly, the point cloud sequences are transformed and voxelized into structured 3D voxel sets. These sets are then augmented with an interval-frequency descriptor to generate 6D features capturing spatiotemporal dynamic information. By evaluating voxel space occupancy using thresholding, we can effectively identify the essential parts. After that, all the voxels with the 6D feature are directed to the global coarse stream, while the voxels within the key parts are routed to the local fine stream. These two streams extract global appearance features and critical body parts by utilizing symmetric PointNet++. Subsequently, attention feature fusion is employed to capture more discriminative motion patterns adaptively. Experiments conducted on public benchmark datasets NTU RGB+D 60 and NTU RGB+D 120 validate SFCNet's effectiveness and superiority for 3D action recognition.

Keywords: point cloud analysis; 3D action recognition; pattern recognition; deep learning

Citation: Li, C.; Huang, Q.; Mao, Y.; Qian, W.; Li, X. A Novel Symmetric Fine-Coarse Neural Network for 3D Human Action Recognition Based on Point Cloud Sequences. *Appl. Sci.* **2024**, *14*, 6335. https://doi.org/10.3390/app14146335

Academic Editor: Atsushi Mase

Received: 11 June 2024
Revised: 8 July 2024
Accepted: 18 July 2024
Published: 20 July 2024

1. Introduction

Human action recognition aims to help computers understand human behavior semantics from various data recorded by the acquisition devices. Particularly, 3D action recognition is dedicated to mining action patterns from 3D data involving human movements. It has attracted increasing attention due to its widespread applications, such as public safety monitoring, performance appraisal, military reconnaissance, and intelligent transportation [1].

The current mainstream 3D action recognition methods can be classified into depth-based methods (including depth maps and point cloud sequences) [2–4] and skeleton-based methods [5,6] depending on the data type employed. Limited by the accuracy pose estimation algorithm—the unavoidable upstream task—skeleton-based methods face computational consumption and robustness challenges. By contrast, depth-based methods are more task-independent and have attracted widespread attention. Existing depth-based 3D action recognition approaches mainly fall into two main categories. The first one is to encode 3D motions into one or more images [2,3,7,8] and utilize CNNs [9] for action recognition. However, the 2D image plane cannot fully characterize the 3D dynamics because human actions are concurrently spatiotemporal and conducted in the 3D space. The other is to transform the depth video into a point cloud sequence [10], which records the 3D coordinates of points in space at multiple time instances. Thus, compared with images, point cloud sequences have the advantage of retaining 3D appearance and

geometry dynamics over time, enabling advanced analysis and understanding of human actions. In addition, point clouds can be obtained using various devices such as laser scanners, radars, depth sensors, and RGB+D cameras, which can be mounted on drones, street lights, vehicles, and surveillance aircraft, expanding the application scope of action recognition. However, due to the complex structure and massive volume of the point cloud, the existing 3D action recognition methods based on it have the following challenges.

Firstly, point cloud sequences always have massive points proportional to the time dimension, and the data processing schema is time-consuming. Therefore, developing an efficient and lightweight point cloud sequence model is pivotal for 3D action recognition. Secondly, the points in the sequences are irregular, exhibiting unordered intra-frame spatial information and ordered inter-frame temporal details, making it challenging to analyze the underlying movement patterns. However, existing point cloud processing methods usually perform undifferentiated downsampling of the overall point clouds, resulting in uniform loss of essential and subtle information. In addition, existing point cloud analysis schemes ignore the critical body parts contributing to the actions, resulting in the lack of nuances of the extracted action features, which finally limits the performance of action recognition.

To solve these problems, we propose a deep learning framework called the Symmetric Fine-coarse Neural Network (SFCNet) that symmetrically combines the analysis of motion features from local and global perspectives, as shown in Figure 1. Firstly, to save computational costs, we reduce the points by frame sampling and farthest point sampling. Next, the sampled point clouds are transformed into 3D voxels to create a compact point cloud representation. The original 3D positions are then attached with an interval-frequency descriptor to depict the overall spatial configuration and facilitate the identification of essential body parts, allowing us to divide the point cloud sequences into local fine space and global coarse space. We treat the voxels involved in these two spaces as points and employ PointNet++ [11] to extract features in an end-to-end manner. Finally, our feature fusion module combines the global appearance and local details to obtain discriminative features for 3D action recognition. The extensive experiments on the large-scale NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate the effectiveness and preponderance of SFCNet, by which human intention can be judged and assisted in remote sensing applications.

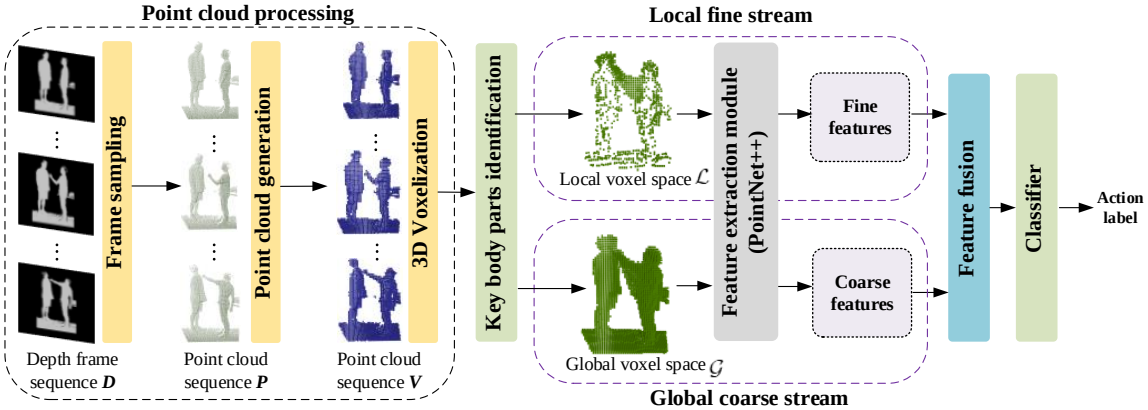


Figure 1. The pipeline of SFCNet. It converts depth frames into a point cloud and applies voxelization operations. A symmetric structure encodes 3D voxels, with crucial parts and global dynamic information processed separately. The attached interval-frequency descriptor initially characterizes the motion information and then is processed by PointNet++ [11] for deeper features. Finally, the classifier recognizes 3D actions using the aggregated feature.

In general, the main contributions of our work are as follows:

- We propose an interval-frequency descriptor to characterize the 3D voxels during action execution, which fully preserves the motion details and provides critical clues for key body parts' perception. To the best of our knowledge, our work is the first to handle point cloud sequences in this way.

- We construct a deep learning framework named SFCNet, which first employs a symmetric structure to process point cloud sequences. It encodes the local crucial body parts' dynamics via a fine stream and then supplements these intricate details to the global appearance captured by a coarse stream. The SFCNet can emphasize essential body parts and capture more discriminative motion patterns, addressing the effective action representation problem based on point clouds.
- The presented SFCNet has demonstrated its superior accuracy on two publicly available datasets, NTU RGB+D 60 and NTU RGB+D 120, which proves that our method has considerable potential in recognizing various types of actions such as daily actions, medical-related actions, and two-person interaction actions.

2. Related Works

2.1. Skeleton-Based 3D Action Recognition

Existing 3D action recognition methods can be classified into skeleton-based methods [12–17] and depth-based methods [3,7,12,18,19]. There are generally four mainstream approaches for skeleton-based action recognition. The first is to utilize CNN [12] to learn the spatial–temporal patterns from pseudo-images [13,14]. Caetano et al. [20] introduced the tree structure reference joints image (TSRJI) to represent skeleton sequences. The second is considering the skeleton sequences as time series [15–17] and using backbones such as RNN [21] for feature extraction. The third is to view the skeletal data as graphs [6,22] with joints as vertices and bones as edges and turn to GCN [16] for action representation. For example, ST-GCN [23] effectively represented the temporal dynamic information of skeleton sequences using spatial–temporal graph convolution and partitioning strategies. SkeleMotion [5] captured temporal dynamic information by computing the size and orientation of skeleton joints at different time scales. The fourth is to encode the skeleton as tokens via Transformer. Plizzari et al. [24] employed spatial–temporal attention in Transformer and captured a dynamic inter-frame relation of joints. However, since there are still significant challenges in accurate 3D human pose estimation [25,26], skeleton-based action recognition methods suffer from performance cascading due to this unavoidable upstream task.

2.2. Depth-Based 3D Action Recognition

For depth-based 3D action recognition, early approaches mainly represent depth videos by manual descriptors [19]. Yang et al. [7] constructed depth motion maps (DMMs) by stacking the inter-frame differences of the projected depth frames. Then, they calculated the histogram of oriented gradients (HOGs) to represent the actions. Such methods have limited expressive power and thus usually need help in capturing spatial–temporal information. In recent years, deep learning methods have become mainstream with the development of neural networks. Most researchers attempted to compress deep video into images and analyzed motion patterns using CNNs [12]. Kamel et al. [27] input depth motion images (DMIs) and moving joint descriptors (MJDs) to CNNs for action recognition. To encode spatial–temporal information of depth sequences, Adrián et al. [28] proposed 3D-CNN to extract motion features. Furthermore, they proposed ConvLSTM [29] to accumulate discriminative motion patterns from long short-term units. Xiao et al. [3] rotated the virtual camera within the 3D space to densely project a raw depth video from different virtual imaging viewpoints and thus constructed multi-view dynamic images. For perspective view invariants, Kumar et al. [30] proposed an ActionNet based on CNNs and trained with a multi-view dataset collected using five depth cameras. Ghosh et al. [31] computed multi-view depth descriptor edge-detected-motion history image (ED-MHI) as the input of a multi-stream CNN model. Wang et al. [2] utilized segmented depth video sequences to generate three types of dynamic depth images. However, the 2D depth map still has difficulty in fully exploiting 3D motion patterns due to its compact spatial structure [10].

Recently, the conversion of depth maps into point clouds for processing has achieved better results in both recognition and segmentation fields. Numerous studies have shown

that point clouds have significant advantages in representing 3D spatial information due to their characteristics, such as disorder and rotation invariance. Deep learning for point clouds has not only been widely used in classification and segmentation tasks but has also demonstrated muscular strength in scene reconstruction [32] and target detection [33]. However, the above methods focus only on features within static point clouds. When using point clouds for 3D action recognition, it is necessary to extract dynamic features according to the time intervals and the appearance features of the whole action process. The key to an efficient processing of point cloud sequences lies in selecting a suitable point cloud analysis method. Thomas et al. [34] developed a method inspired by image-based convolution and employed a set of kernel points to distribute each kernel weight. As an efficient tool for analyzing and processing point sets, PointNet++ [11] is widely applied for 3D action recognition based on point cloud sequences. The first method is 3DV [10], which executes 3D voxelization towards the point cloud sequences and describes the 3D appearance by spatial occupancy, and temporal rank pooling is utilized for 3DV extraction. This method primarily focuses on an action's general motion and appearance changes. However, it ignores the details of the action, such as a subtle hand movement, which limits its ability to represent behavior accurately. Hence, we aim to capture the crucial parts of actions and their delicate information to recognize them as more robust human actions.

3. Methodology

3.1. Pipeline

The pipeline of the proposed SFCNet is shown in Figure 1. First, each depth frame is transformed into a point cloud to better preserve the dynamic and appearance features in the 3D space. To facilitate the analysis of spatial usage and delineate the local space, we perform voxelization operations on the point clouds. Next, we build a symmetric framework to encode 3D voxels, where key parts and global dynamic information are processed separately in the local fine stream and global coarse stream. Then, we attach the interval-frequency descriptor to supplement motion information. We employ PointNet++ [11] to capture motion patterns and send the aggregated feature to the classifier for 3D action recognition.

3.2. Three-Dimensional Voxel Generation

Depth video has the advantage of resisting external interference, such as background and light, compared with RGB modal because it contains the depth information of the action subject. Essentially, depth video is a kind of time series data composed of depth maps arranged in chronological order. Mathematically, a depth video with t frames can be defined as $D = \{d_1, d_2, \ldots, d_t\}$, where d_i is a depth map of t frame in which each pixel represents a 3D coordinate (x, y, z) and z is the distance from the depth camera. Since it is impossible to classify the importance of action in the time dimension by a single criterion, uniform sampling can help us better grasp the overall motion process compared with random sampling [10]. Therefore, we first sample the depth video uniformly to ease the computational burden while maintaining the integrity of the action. The depth sequence after sampling is denoted as $\hat{D} = \{d_1, d_2, \ldots, d_T\}$, where T is the number of frames and the default is 64.

Some current action recognition methods [35,36] choose to map depth frames to 2D spaces for direct processing. Although these approaches can sometimes achieve good performance, it cannot overcome the problem of inadequate representation of 3D information. Therefore, to better depict human motion in 3D space, we transform each frame d_i into a point cloud $P = \{p_1, p_2, \ldots, p_n\}$, where n is the number of points, thus generating a point cloud sequence $S = \{P_1, P_2, \ldots, P_T\}$. When generating point clouds, intrinsic parameters of the camera are required because they define the imaging model of the camera, including

focal length and principal point coordinates (c_x, c_y). For each pixel (x, y, z) in the depth image, its corresponding point cloud $p(x^{'}, y^{'}, z^{'})$ can be obtained by the following formula:

$$p(x^{'}, y^{'}, z^{'}) = (\frac{(x - c_x) \times z}{f_x}, \frac{(y - c_y) \times z}{f_y}, \frac{z}{f_z}) \quad (1)$$

where f_x and f_y represent the focal length of the depth camera in the horizontal direction and the vertical direction, which can be obtained from the device parameters. f_z is set to 1 by default.

Unlike traditional images (regular structured data), the points in the point cloud are unordered, so it is challenging to process. Many existing algorithms are designed for regular grid data. However, the unordered point cloud is a group of randomly distributed points in 3D space, so their structure is complex to process and analyze directly. To solve this problem, we transform the point cloud into a regular 3D grid (voxel space) by voxelization to regularize the point cloud representation. First, we define the size of voxel grid $V_{grid} = (V_x, V_y, V_z)$ in three dimensional coordinates, which determines the resolution of the voxelization process. Each cell in this grid is a potential voxel and the size of each cell is denoted as $V_{voxel}(d_x, d_y, d_z)$. Given a point $p(x, y, z)$ in the point cloud, it is mapped to the grid by finding the corresponding voxel index $V_{index}(x, y, z)$ according to the following equation:

$$V_{index}(x, y, z) = (\lfloor \frac{x - x_{min}}{d_x} \rfloor, \lfloor \frac{y - y_{min}}{d_y} \rfloor, \lfloor \frac{z - z_{min}}{d_z} \rfloor) \quad (2)$$

where x_{min}, y_{min}, and z_{min} are the minimum coordinates of all the point clouds. d_x, d_y, and d_z are calculated as the total size divided by the number of cells in each dimension (V_x, V_y, V_z). The floor function $\lfloor . \rfloor$ rounds down to the nearest point. We define that a voxel is occupied if it contains a point cloud. Then, the 3D appearance information can be described by observing whether the voxels have been occupied or not, disregarding the excluded point, as depicted in Equation (3):

$$V^t_{voxel}(x, y, z) = \begin{cases} 1, & \text{if } V^t_{voxel}(x, y, z) \text{ is occupied} \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

where $V^t_{voxel}(x, y, z)$ indicates a certain voxel at the t_{th} frame. (x, y, z) is the regular 3D position index, i.e., V_{index} in Equation (2). This strategy holds two main profits. First, the yielded binary 3D voxel sets are regular, as depicted in Figure 2. Thus, the complexity of point cloud processing is reduced. In addition, voxelization can effectively compress point clouds because neighboring voxels may have similar characteristics. This compression not only reduces the amount of points but also helps to reduce the overhead of storage and computation.

3.3. Identification and Representation of Key Parts

The vital issue in 3D action recognition tasks is efficiently capturing and representing dynamic features within point cloud sequences. For now, estimation methods based on scene flow [37,38] can help to understand 3D motion, but it is very time-consuming. Some studies use temporal rank pooling [3,39] to preserve motion processes in 3D space by dividing time segments. These methods can capture more temporal information but often only divide a small number of intervals, resulting in coarse-grained dynamic features. We propose a crucial part identification and encoding module to better focus on the critical dynamics during the motion. It can extract the major parts from the global 3D voxel space according to the space occupancy times and encode the action details through fine-stream. Specifically, we first analyze the space occupancy by constructing a 3D space U with the exact boundaries as point cloud sequences for each spatial location, and the initial values

of U are set to 0. We process the m uniform groups of sequence $\hat{D}$ in order. Then, the 3D space usage can be calculated according to Equation (4):

$$U_{voxel}(x,y,z) = \begin{cases} U_{voxel}(x,y,z)+1, v_i \neq 0 \\ U_{voxel}(x,y,z), otherwise \end{cases} \tag{4}$$

The total space occupation u for each position can be obtained after counting all m point sets. In addition, since the point sets are naturally ordered in time, we can easily record the first and last time taken, f and l, respectively, for each spatial location. Then, we define the threshold θ to partition the prominent local space. The locations occupied less than θ are treated as incidental noise, and those that record more than and less than m constitute the critical motion parts S as Equation (5):

$$S_{voxel}(x,y,z) = \begin{cases} 0, & U_{voxel}(x,y,z) < \theta \\ 1, & \theta \leq U_{voxel}(x,y,z) \leq m \end{cases} \tag{5}$$

If the value of $S_{voxel}(x,y,z)$ equals 0, this denotes that the voxel belongs to the global space $\mathcal{G}$; otherwise, it belongs to the local space $\mathcal{L}$. Compared with the point cloud processing methods [10,40], which commonly adopt uniform downsampling operations, the proposed method is more effective, especially for actions involving only a tiny number of limb parts, because dividing the local space can not only also overcome the background effects to a certain extent, but also effectively enhance the gold content of the sampled point data. As shown in Figure 2, the local space $\mathcal{L}$ fully preserves the detailed information of the main body parts, which provides critical cues for 3D action recognition while substantially reducing redundancy.

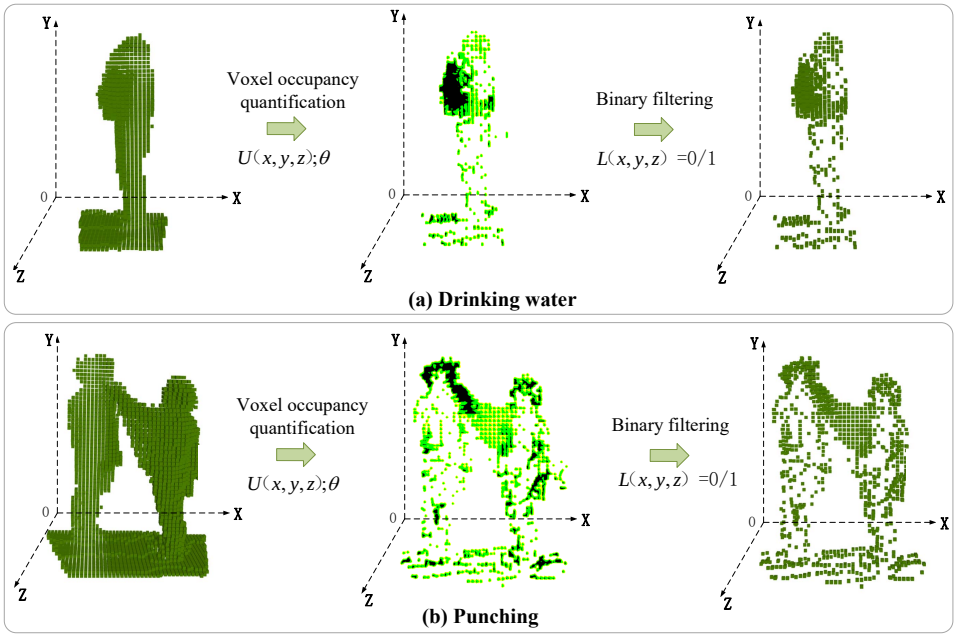


Figure 2. The process of local space division. We quantify the contribution of a voxel to an action by its space occupancy count. By setting a threshold, the critical parts can be divided as a compressed local space, which removes redundant information and reduces computational burden.

3.4. Symmetric Feature Extraction

For the processed 3D voxels, the most intuitive way is to employ 3DCNN [41,42], but it is limited by the voxel size and is time-consuming. We opt to use PointNet++ [11] as the feature extractor in this work, as shown in Figure 3. It is designed explicitly for hierarchical feature learning on unordered point sets in metric space, which allows it to capture local fine-grained patterns in point clouds. To achieve this, PointNet++ partitions the point

cloud into overlapping local regions based on a distance metric in the underlying space. To obtain detailed 3D visual cues, PointNet++ recursively employs PointNet [43] to extract local features, which are then merged for global appearance analysis. PointNet++ is an excellent alternative to 3DCNNs as it performs well in capturing local 3D patterns essential for action recognition. Moreover, applying it is relatively straightforward and only requires transforming 3D voxels into point sets.

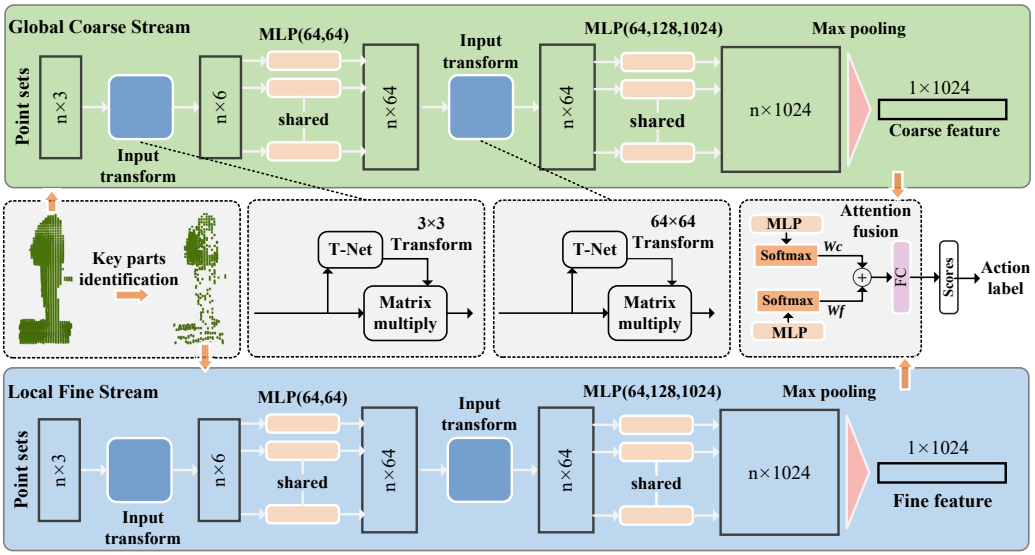


Figure 3. The network structure of SFCNet. It is a symmetric network structure comprising the global coarse stream and a local fine stream. The global coarse stream takes the global appearance information as input, while the local fine stream only adopts the compressed key part information. The features of each stream are extracted by PointNet++ and fused by the learnable weights as the final action features, which are sent to the classifier for 3D action recognition.

To fit PointNet++, each point $p(x, y, z)$ is then abstracted as the voxel $V_{voxel}(x, y, z)$ with the descriptive feature of (x, y, z, I), where I is the interval-frequency descriptor including three variables (o, f, l) which denote start timestamps, end timestamps, and the overall occupancy frequency of the voxels, respectively. We attach I to the original 3D positions of voxels to obtain the 6D points $p^{'}(x, y, z, o, f, l)$, as shown in Figure 4. Mathematically, for a voxel V_{voxel}, we can determine the start and end time (o and f) of its space occupancy using Equation (3), which indicates the time index when the conditions $V^{t}_{voxel}(x, y, z) = 1$ and $V^{t}_{voxel}(x, y, z) = 0$ are first satisfied. Additionally, the total space occupation u can be calculated by Equation (4). Consequently, the occupancy frequency l can be computed as $l = u/(f - o)$. The interval-frequency descriptor covers time interval and the overall occupancy frequency of the spatial locations can not only help us to distinguish the reverse actions that cover the same space, but also help to highlight the differences between actions by retaining detailed information within action process. Finally, the obtained point sets are utilized as the input of PointNet++ to extract action features.

In addition, we design a two-stream network to process the point cloud in the global and local space, respectively (as shown in Figure 3). The global space contains all the voxelized points, and the global coarse stream captures the overall motion patterns. Considering that vital body parts can provide more targeted and discriminative dynamic information for action recognition, we divide the voxelized points of essential body parts into local spaces and input the fine stream to extract features. After that, the feature fusion module is settled for fine-coarse action representation. Considering the characteristics of different actions, their dependence on global and local features is various. For actions that only involve limb movements, such as waving and kicking, the model should emphasize local fine-grained stream features. In contrast, for large whole-body motions such as falling and jumping, the model should focus on the characteristics of the global stream. To achieve

this action-specific perception, we employ the feature fusion module with attention mechanism. First, we project the features $X_f \in R^{n\times 1024}$ and $X_c \in R^{n\times 1024}$ extracted from the fine and coarse streams, respectively, into the lower feature space to reduce the computational burden and obtain $X_f^{'}$ and $X_c^{'}$. Then, the intermediate features are extracted by multi-layer perceptron (MLP). After that, the learnable weights W_f and W_c of global coarse stream and local fine stream are obtained through the activation function SoftMax, respectively. Finally, the global and local features are fused as shown in Equation (6) to obtain the motion feature $\hat{X}$:

$$\begin{aligned} W_f &= \mathrm{SoftMax}\left(\mathrm{MLP}\left(X_f'\right)\right) \\ W_c &= \mathrm{SoftMax}\left(\mathrm{MLP}\left(X_c'\right)\right) \\ \hat{X} &= \varphi\left(\sum_{i=1}^{K}\left(W_f^i X_f^{(i)} + W_c^i X_c'^{(i)}\right)\right) \end{aligned} \tag{6}$$

where φ is the linear layer and K is the length of the features' output by MLP. Finally, the fully connected layer dimensionally restored the fused feature $\hat{X}$, and the SoftMax classifier obtained the final prediction scores. Unlike existing methods that directly analyze the motion state of the whole point clouds, our work focuses on the critical body part when performing actions, which helps to overcome the influence of redundant data such as background on 3D action recognition. In addition, the details of the crucial parts and the global appearance of human body complement each other, which stimulates discriminative feature extraction for 3D action recognition.

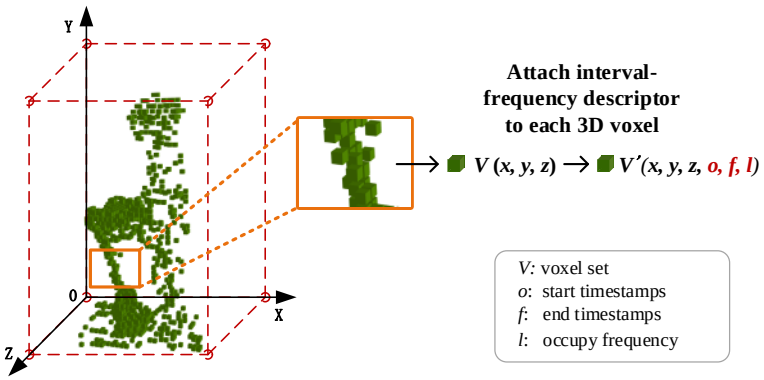


Figure 4. Illustration of the additional interval-frequency descriptor. It contains start timestamps, end timestamps, and the overall occupancy frequency of the voxels and is denoted as (o, f, l) in Section 3.4; thus, the 3D depth spatial information is transformed into features of six channels.

4. Experiments

4.1. Datasets

NTU RGB+D 60 dataset . The NTU RGB+D 60 [44] is a large-scale 3D action recognition dataset that contains around 56,880 RGB+D action samples. It uses Microsoft Kinect V2 to capture 60 action categories performed by 40 subjects. The dataset follows two evaluation principles. In the case of cross-view, the samples captured by camera 1 were used as the testing set, and cameras 2 and 3 are regarded as the training set. That is, the number of testing samples is 18,960, and 37,920 samples are used for training. In the case of cross-subject, the data are divided into a testing set of 16,560 samples and a training set of 40,320 samples based on the subject's ID.

NTU RGB+D 120 dataset. NTU RGB+D 120 [45] is an extensive dataset for 3D action recognition, consisting of 114,480 samples and 120 action categories completed by 106 subjects. This dataset contains daily actions, medical-related actions, and two-person interaction actions. The samples are collected in various locations and backgrounds, which are denoted as 32 setups. In addition to the general cross-subject settings, the cross-setup

evaluation is introduced, where the training set comes from samples with odd setup IDs, and the testing set comes from the rest.

4.2. Training Details

By default, the SFCNet and its variations are trained using the Adaptive Moment Estimation (Adam) optimizer for 60 epochs under the PyTorch deep learning framework, unless stated otherwise. We use the standard cross-entropy loss and apply data augmentation techniques such as random rotation, dithering, and dropout to the training data. The learning rate starts from 0.001 and decays at 0.5 every ten epochs. To ensure fairness, we strictly follow the sample segmentation scheme of the two datasets according to benchmarks.

4.3. Parameter Analysis

The size of 3D voxels. Point clouds typically consist of a large number of unordered sets of points. Due to the complexity of these data, it can be very time-consuming to store and process. To alleviate this issue, we embed the unordered points in 3D space into a regular grid structure through rasterization. This converts the point cloud into 3D voxels based on space occupancy, discretizing the continuous 3D space into a regular grid. This process provides a regular and well-understood structure, which reduces computational complexity and compresses the point cloud data, significantly reducing computational burden. It is essential to voxelize the point cloud appropriately because the size of the 3D voxel determines the strength of the point cloud compression and the granularity of the point cloud representation. To examine the impact of voxelization on the results, we evaluated the performance of SFCNet on the NTU RGB+D 60 dataset for different-sized voxels. The results are presented in Table 1, indicating that the model performs best for a 35 mm cube size. Setting the size too large or too small can lead to a decrease in accuracy.

Table 1. Performance on the NTU RGB+D 60 dataset with different-sized voxelization.

Voxel Size (mm)	Cross-Subject	Cross-View
$25 \times 25 \times 25$	87.1%	94.9%
$35 \times 35 \times 35$	**89.9%**	**96.7%**
$45 \times 45 \times 45$	88.1%	95.5%
$55 \times 55 \times 55$	86.5%	93.6%

The setting of threshold θ. Human behavior usually involves only the movement of specific body parts, such as arm waving, leg walking, head rotation, etc. This locality means that behavioral analysis should be more focused on essential body parts rather than the whole body. With the help of the occupancy frequency variable l in the interval-frequency descriptor, we can describe the engagement of each voxel, which is positively related to the contribution of the body part in the action execution process. Since we sample the depth action sequence into groups of equal duration, the occupancy frequency l positively correlates with the number of occupancy u in Section 3.3. Then, a threshold θ is employed to evaluate the attention given to the body part. To investigate the influence of threshold, we compare the performance of the SFCNet on the NTU RGB+D 60 dataset with various values. The findings are presented in Table 2. The optimal outcome can be obtained when θ equals 30. Our research discovered that slight modifications in θ, by no more than 5, resulted in fluctuations in accuracy, underscoring the significance of investigating thresholds and breaking down significant body parts.

Table 2. Performance on the NTU RGB+D 60 dataset with various values of θ.

The Value of the Threshold θ	Cross-Subject	Cross-View
15	79.5%	85.1%
20	83.5%	93.2%
25	86.5%	94.4%
30	**89.9%**	**96.7%**
35	87.3%	94.9%
40	86.9%	93.7%

4.4. Ablation Study

Effectiveness of interval-frequency descriptor. The original 3D point cloud data only contain the location information of the points in the 3D space. Even if the time dimension is introduced into the point cloud sequences, it is still challenging to describe the overall spatiotemporal dynamics of the point by relying only on these clues. We have designed an interval-frequency descriptor that captures the onset time and the number of voxel occupancy. This additional information helps us to comprehensively describe human behavior by capturing additional motion features. We conducted ablation studies on the NTU RGB+D 60 dataset where we removed motion feature information in two streams, and the point sets input to SFCNet only had 3D coordinates (x, y, z). The comparison results are presented in Table 3. We observed that without additional three-dimensional features, there was a significant performance degradation of SFCNet by more than 10%.

Table 3. Effectiveness of interval-frequency descriptor on the NTU RGB+D 60.

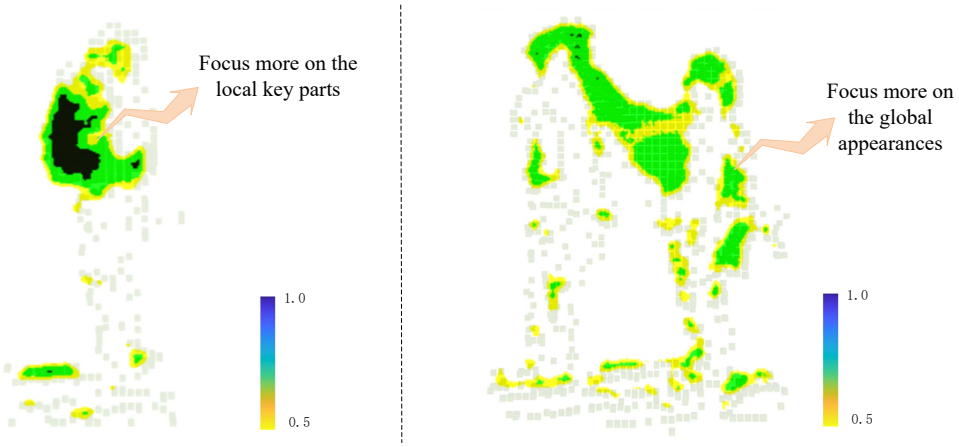
Point Feature	Cross-Subject	Cross-View
(x, y, z)	78.0%	82.3%
(x, y, z, o, f, l)	**89.9%**	**96.7%**

This indicates that the interval-frequency descriptor effectively represents the dynamic features within the whole action process, which plays a vital role in 3D action recognition.

Effectiveness of two-stream feature fusion. Different human actions contain different global and local dynamics, we describe the overall motion pattern of the whole human body during the action through the global coarse stream. In contrast, the local fine stream describes the dynamics of crucial body parts, which pays more attention to the details and local features of actions and helps capture the subtle changes and complex action patterns. The results in Table 4 show that our proposed SFCNet fusing the fine-coarse features of the two streams can understand and recognize human actions more comprehensively. First, we discuss the recognition performance in the single-stream state. It can be seen that the local stream has a higher ability to represent the action than the global stream, indicating that it is essential to pay attention to the main parts involved to remove redundancy. In addition, we compare three different feature fusion strategies to demonstrate the superiority of attention-based feature fusion proposed in SFCNet (see Equation (6)). As shown in Table 4, SFCNet (fusion) has apparent advantages over the native cascade or additive fusion strategies. The main reason is that the attention-based feature fusion can adaptively allocate the model's attention to the features from the global coarse stream and the local fine stream. As shown in Figure 5, drinking water only involves hand and head interaction, and the movement amplitude is small, so the model emphasizes the characteristics of the local stream to capture fine-grained movement patterns. On the other hand, punching involves the interaction of two people, and the punching movement is large and powerful, so more attention is paid to the global body appearance while emphasizing some details of the hands and the head. This action-specific feature extraction mechanism improves the generalization ability of SFCNet and accuracy for 3D action recognition.

Table 4. Effectiveness of two-stream feature fusion on the NTU RGB+D 60 dataset.

Input Stream	Cross-Subject	Cross-View
1s-SFCNet (L)	85.0%	94.6%
1s-SFCNet (G)	81.8%	86.6%
SFCNet (concat)	88.9%	94.8%
SFCNet (add)	86.7%	93.9%
SFCNet (fusion)	**89.9**%	**96.7**%



(a) Drinking water **(b) Punching**

Figure 5. Feature attention visualization. We visualize the heat map of drinking water and punching.

4.5. Comparison with Existing Methods

To evaluate the performance of the proposed SFCNet, we compare it with existing methods on two large benchmark datasets as shown in Tables 5 and 6. We divide existing 3D action recognition methods into skeleton-based and depth-based methods. In skeleton-based methods, we compare different backbone-based methods, including CNN [5], LSTM [15,46], and GCN [6,23,47]. For depth-based methods, we compare 2D image-based methods [19,35,48], 3D CNN-based methods [28], and 3D voxel-based methods [10]. The comparison results are reported in Tables 5 and 6. For the NTU RGB+D 60 dataset, the proposed SFCNet achieves 89.9% and 96.7% accuracy in the case of cross-subject and cross-view settings. In addition, we compare SFCNet with two methods based on multimodal data. Compared with ED-MHI [31], which combined depth and skeleton data, our method improves the accuracy by 4.3% in the cross-subject setting. TS-CNN-LSTM [49] fused data from three modalities, namely RGB, depth, and skeleton, but it is 2.6% and 4.9% lower than SFCNet in the cross-subject setting and cross-view settings, respectively. For the NTU RGB+D 120 dataset, SFCNet also achieves competitive results, achieving accuracies of 83.6% and 93.8% under cross-subject and cross-view settings, respectively. In general, SFCNet is effective and excellent, which outperforms traditional methods using manual feature extraction [19,35,50] and deep learning methods that compress depth video into images for processing [2,3,36] or point cloud sequences [10]. The experimental results prove that the SFCNet is superior for capturing discriminative human behavior patterns and thus is beneficial to 3D action recognition.

Table 5. Comparison of different methods for action recognition accuracy (%) on the NTU RGB+D 60 dataset.

Method	Cross-Subject	Cross-View	Year
Input: 3D Skeleton			
GCA-LSTM [15]	74.4	82.8	2017
Two-stream attention LSTM [46]	77.1	85.1	2018
ST-GCN [23]	81.5	88.3	2018
SkeleMotion [5]	69.6	80.1	2019
AS-GCN [6]	86.8	94.2	2019
2s-AGCN [47]	88.5	95.1	2019
ST-TR (new) [24]	89.9	96.1	2021
DSwarm-Net (new) [51]	85.5	90.0	2022
ActionNet [30]	73.2	76.1	2023
SGMSN (new) [52]	90.1	95.8	2023
Input: Depth maps			
HON4D[19]	30.6	7.3	2013
HOG^2 [35]	32.2	22.3	2013
SNV [50]	31.8	13.6	2014
Li. [36]	68.1	83.4	2018
Wang. [2]	87.1	84.2	2018
MVDI [3]	84.6	87.3	2019
3DV-PointNet++ [10]	88.8	96.3	2020
DOGV (new) [53]	90.6	94.7	2021
3DFCNN [28]	78.1	80.4	2022
3D-Pruning [54]	83.6	92.4	2022
ConvLSTM (new) [29]	80.4	79.9	2022
CBBMC (new) [48]	83.3	87.7	2023
PointMapNet (new) [55]	89.4	96.7	2023
SFCNet (ours)	**89.9**	**96.7**	-
Input: Multimodalities			
ED-MHI [31]	85.6	-	2022
TS-CNN-LSTM [49]	87.3	91.8	2023

Table 6. Comparison of different methods for action recognition accuracy (%) on the NTU RGB+D 120 dataset.

Method	Cross-Subject	Cross-Set	Year
Input: 3D Skeleton			
GCA-LSTM [15]	58.3	59.3	2017
Body pose evolution map [56]	64.6	66.9	2018
Two-stream attention LSTM [46]	61.2	63.3	2018
ST-GCN [23]	70.7	73.2	2018
NTU RGB+D 120 baseline [45]	55.7	57.9	2019
FSNet [57]	59.9	62.4	2019
SkeleMotion [5]	67.7	66.9	2019
TSRJI [20]	67.9	62.8	2019
AS-GCN [6]	77.9	78.5	2019
2s-AGCN [47]	82.9	84.9	2019
ST-TR (new) [24]	82.7	84.7	2021
SGMSN (new) [52]	84.8	85.9	2023
Input: Depth maps			
APSR [45]	48.7	40.1	2019
3DV-PointNet++ [10]	82.4	93.5	2020
DOGV (new) [53]	82.2	85.0	2021
3D-Pruning [54]	76.6	88.8	2022
SFCNet (ours)	**83.6**	**93.8**	-

5. Discussion

In order to analyze the advantages and disadvantages of the proposed method, we have presented the recognition accuracy of SFCNet on the NTU RGB+60 dataset on the cross-subject settings for each category. The results are shown in the form of a confusion matrix in Figure 6(left). We have selected some confusing actions for a clearer display and magnified them locally, as shown in Figure 6(right). The results show that SFCNet has a robust human action analysis ability, with a recognition accuracy higher than 90% in most categories. For instance, it has achieved 100% accuracy for hopping and 99% for jumping up. However, SFCNet is confused about the recognition of some similar actions. For example, reading and writing, wearing shoes and taking off shoes are the most confusing sample pairs. Additionally, 25% of those playing with their phones were misclassified as reading (9%), writing (8%), and typing on the keyboard (8%). The accuracy of typing on the keyboard is only 66%, and 12% of the samples are misclassified as writing. From the analysis, we have found that these actions have only subtle differences, and the motion amplitude is small. This is the main reason why such actions are difficult to distinguish.

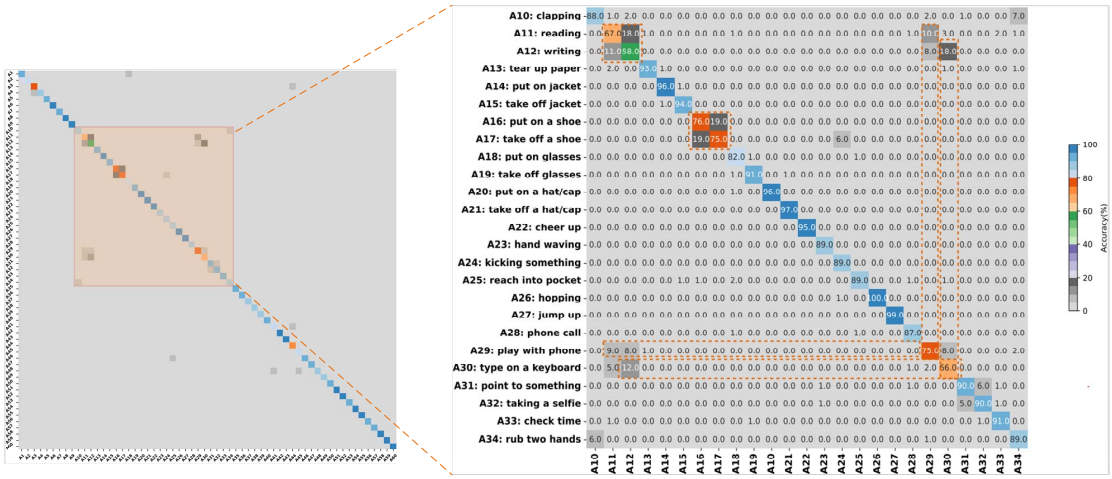


Figure 6. Confusion matrix for class-specific recognition accuracy. The image on the (**left**) contains all the action categories. To emphasize some of the obfuscation actions, a local zoom-in is shown on the (**right**).

6. Conclusions

In this paper, we propose a symmetric neural network, SFCNet, to recognize 3D actions from point cloud sequences. It contains a global coarse stream and a local fine stream that employs PointNet++ as a feature extractor. The point cloud sequences are regularized as structured voxel sets appended by the proposed interval-frequency descriptor to generate 6D features that capture spatiotemporal dynamic information. The global coarse stream captures the coarse-grained action patterns by human body appearance, and the local delicate stream extracts action-specific fine-grained features from critical parts. After feature fusion, SFCNet can mine discriminative motion patterns that involve overall spatial changes and emphasize crucial details end-to-end. According to the experimental results on two large benchmark datasets, NTU RGB+D 60 and NTU RGB+D 120, the SFCNet is effective for 3D action recognition and has the potential for remote sensing applications. However, the proposed SFCNet still has limitations in distinguishing similar actions. Our future work will focus on recognizing similar actions and capturing subtle patterns to improve accuracy.

Author Contributions: Conceptualization, C.L. and Q.H.; methodology, C.L., W.Q. and X.L.; software, Q.H. and Y.M.; validation, C.L., W.Q. and X.L.; formal analysis, Q.H. and Y.M.; investigation, C.L. and W.Q.; resources, Q.H. and Y.M.; data curation, W.Q.; writing—original draft preparation, C.L. and W.Q.; writing—review and editing, C.L., Y.M., W.Q., Q.H. and X.L.; visualization, W.Q.; supervision,

Q.H. and Y.M.; project administration, Q.H. and Y.M.; funding acquisition, Q.H., Y.M. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant number KYCX23_0753), the Fundamental Research Funds for the Central Universities (grant number B230205027), the Key Research and Development Program of China (grant number 2022YFC3005401), the Key Research and Development Program of China, Yunnan Province (grant number 202203AA080009), the 14th Five-Year Plan for Educational Science of Jiangsu Province (grant number D/2021/01/39), and the Jiangsu Higher Education Reform Research Project (grant number 2021JSJG143); and the APC was funded by the Fundamental Research Funds for the Central Universities.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The NTU RGB+D 60 and NTU RGB+D 120 datasets used in this paper are public, free, and available at: https://rose1.ntu.edu.sg/dataset/actionRecognition/ (accessed on 21 December 2020).

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

References

1. Riaz, W.; Gao, C.; Azeem, A.; Saifullah; Bux, J.A.; Ullah, A. Traffic Anomaly Prediction System Using Predictive Network. *Remote Sens.* **2022**, *14*, 1–19.
2. 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.
3. Xiao, Y.; Chen, J.; Wang, Y.; Cao, Z.; Zhou, J.T.; Bai, X. Action recognition for depth video using multi-view dynamic images. *Inf. Sci.* **2019**, *480*, 287–304.
4. Li, C.; Huang, Q.; Li, X.; Wu, Q. Human action recognition based on multi-scale feature maps from depth video sequences. *Multimed. Tools Appl.* **2021**, *80*, 32111–32130.
5. Caetano, C.; Sena, J.; Brémond, F.; Dos Santos, J.A.; Schwartz, W.R. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, Taipei, Taiwan, 18–21 September 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
6. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3595–3603.
7. Yang, X.; Zhang, C.; Tian, Y. Recognizing actions using depth motion maps-based histograms of oriented gradients. In Proceedings of the ACM International Conference on Multimedia, Nara, Japan, 29 October–2 November 2012; pp. 1057–1060.
8. 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.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
10. 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. 508–517.
11. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5105–5114.
12. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P.O. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Hum. -Mach. Syst.* **2015**, *46*, 498–509.
13. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3288–3297.
14. Li, C.; Zhong, Q.; Xie, D.; Pu, S. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv* **2018**, arXiv:1804.06055v1 .
15. Liu, J.; Gang, W.; Ping, H.; Duan, L.Y.; Kot, A.C. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3671–3680.
16. 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.

17. Si, C.; Chen, W.; Wang, W.; Wang, L.; Tan, T. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1227–1236.
18. Yu, Z.; Wenbin, C.; Guodong, G. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image Vis. Comput.* **2014**, *32*, 453–464.
19. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723.
20. Caetano, C.; Brémond, F.; Schwartz, W.R. Skeleton Image Representation for 3D Action Recognition Based on Tree Structure and Reference Joints. In Proceedings of the Thirty-second SIBGRAPI Conference on Graphics, Patterns and Images, Rio de Janeiro, Brazil, 28–31 October 2019; pp. 16–23.
21. 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 2022; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
22. 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.
23. 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; pp. 7444–7452.
24. Plizzari, C.; Cannici, M.; Matteucci, M. Skeleton-based action recognition via spatial and temporal transformer networks. *Comput. Vis. Image Underst.* **2021**, *208-209*, 103219.
25. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-time human pose recognition in parts from single depth images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1297–1304.
26. Xiong, F.; Zhang, B.; Xiao, Y.; Cao, Z.; Yu, T.; Zhou, J.T.; Yuan, J. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 793–802.
27. 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.
28. Sánchez-Caballero, A.; de López-Diz, S.; Fuentes-Jimenez, D.; Losada-Gutiérrez, C.; Marrón-Romera, M.; Casillas-Pérez, 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.
29. Sánchez-Caballero, A.; Fuentes-Jiménez, D.; Losada-Gutiérrez, C. Real-time human action recognition using raw depth video-based recurrent neural networks. *Multimed. Tools Appl.* **2022**, *82*, 16213–16235.
30. Kumar, D.A.; Kishore, P.V.V.; Murthy, G.; Chaitanya, T.R.; Subhani, S. View Invariant Human Action Recognition using Surface Maps via convolutional networks. In Proceedings of the International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering, Chennai, India, 1–2 November 2023; pp. 1–5.
31. Ghosh, S.K.; M, R.; Mohan, B.R.; Guddeti, R.M.R. Deep learning-based multi-view 3D-human action recognition using skeleton and depth data. *Multimed. Tools Appl.* **2022**, *82*, 19829–19851.
32. Li, R.; Li, X.; Fu, C.W.; Cohen-Or, D.; Heng, P.A. Pu-gan: a point cloud upsampling adversarial network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7203–7212.
33. Qi, C.R.; Litany, O.; He, K.; Guibas, L.J. Deep hough voting for 3d object detection in point clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9277–9286.
34. Thomas, H.; Qi, C.R.; Deschaud, J.E.; Marcotegui, B.; Goulette, F.; Guibas, L. KPConv: Flexible and Deformable Convolution for Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6410–6419.
35. Ohn-Bar, E.; Trivedi, M.M. Joint Angles Similarities and HOG2 for Action Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.
36. Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Unsupervised learning of view-invariant action representations. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1262–1272.
37. Liu, X.; Qi, C.R.; Guibas, L.J. Flownet3d: Learning scene flow in 3d point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 529–537.
38. Zhai, M.; Xiang, X.; Lv, N.; Kong, X. Optical flow and scene flow estimation: A survey. *Pattern Recognit.* **2021**, *114*, 107861.
39. Fernando, B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 773–787.
40. Liu, J.; Xu, D. GeometryMotion-Net: A strong two-stream baseline for 3D action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4711–4721.
41. Dou, W.; Chin, W.H.; Kubota, N. Growing Memory Network with Random Weight 3DCNN for Continuous Human Action Recognition. In Proceedings of the IEEE International Conference on Fuzzy Systems, Incheon, Republic of Korea, 13–17 August 2023; pp. 1–6.

42. Fan, H.; Yu, X.; Ding, Y.; Yang, Y.; Kankanhalli, M. PSTNet: Point spatio-temporal convolution on point cloud sequences. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 Arpil 2020; pp. 1–6.
43. 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.
44. 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/CVF Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
45. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701.
46. Liu, J.; Wang, G.; Duan, L.Y.; Abdiyeva, K.; Kot, A.C. Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks. *IEEE Trans. Image Process.* **2018**, *27*, 1586–1599.
47. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12026–12035.
48. Li, X.; Huang, Q.; Wang, Z. Spatial and temporal information fusion for human action recognition via Center Boundary Balancing Multimodal Classifier. *J. Vis. Commun. Image Represent.* **2023**, *90*, 103716.
49. Zan, H.; Zhao, G. Human Action Recognition Research Based on Fusion TS-CNN and LSTM Networks. *Arab. J. Sci. Eng.* **2023**, *48*, 2331–2345.
50. Yang, X.; Tian, Y. Super normal vector for activity recognition using depth sequences. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 804–811.
51. Basak, H.; Kundu, R.; Singh, P.K.; Ijaz, M.F.; Woźniak, M.; Sarkar, R. A union of deep learning and swarm-based optimization for 3D human action recognition. *Sci. Rep.* **2022**, *12*, 1–17.
52. Qi, Y.; Hu, J.; Zhuang, L.; Pei, X. Semantic-guided multi-scale human skeleton action recognition. *Appl. Intell. Int. J. Artif. Intell. Neural Netw. Complex Probl. -Solving Technol.* **2023**, *53*, 9763–9778.
53. Ji, X.; Zhao, Q.; Cheng, J.; Ma, C. Exploiting spatio-temporal representation for 3D human action recognition from depth map sequences. *Knowl. -Based Syst.* **2021**, *227*, 107040.
54. Guo, J.; Liu, J.; Xu, D. 3D-Pruning: A Model Compression Framework for Efficient 3D Action Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 8717–8729.
55. Li, X.; Huang, Q.; Zhang, Y.; Yang, T.; Wang, Z. PointMapNet: Point Cloud Feature Map Network for 3D Human Action Recognition. *Symmetry* **2023**, *15*, 1–17.
56. Liu, M.; Yuan, J. Recognizing human actions as the evolution of pose estimation maps. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1159–1168.
57. Liu, J.; Shahroudy, A.; Wang, G.; Duan, L.Y.; Kot, A.C. Skeleton-based online action prediction using scale selection network. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 1453–1467.