



Article

3D-STARNET: Spatial–Temporal Attention Residual Network for Robust Action Recognition

Jun Yang ^{1,2} , Shulong Sun ^{2,*} , Jiayue Chen ¹, Haizhen Xie ¹, Yan Wang ¹ and Zenglong Yang ¹

¹ Big Data and Internet of Things Research Center, China University of Mining and Technology, Beijing 100083, China; yj@cumtb.edu.cn (J.Y.); sqt2200405085@student.cumtb.edu.cn (J.C.); zqt2200405142@student.cumtb.edu.cn (H.X.); sqt2200405099@student.cumtb.edu.cn (Y.W.); zqt2310405043@student.cumtb.edu.cn (Z.Y.)

² Key Laboratory of Intelligent Mining and Robotics, Ministry of Emergency Management, Beijing 100083, China

* Correspondence: sslong@student.cumtb.edu.cn

Abstract: Existing skeleton-based action recognition methods face the challenges of insufficient spatiotemporal feature mining and a low efficiency of information transmission. To solve these problems, this paper proposes a model called the Spatial–Temporal Attention Residual Network for 3D human action recognition (3D-STARNET). This model significantly improves the performance of action recognition through the following three main innovations: (1) the conversion from skeleton points to heat maps. Using Gaussian transform to convert skeleton point data into heat maps effectively reduces the model’s strong dependence on the original skeleton point data and enhances the stability and robustness of the data; (2) a spatiotemporal attention mechanism (STA). A novel spatiotemporal attention mechanism is proposed, focusing on the extraction of key frames and key areas within frames, which significantly enhances the model’s ability to identify behavioral patterns; (3) a multi-stage residual structure (MS-Residual). The introduction of a multi-stage residual structure improves the efficiency of data transmission in the network, solves the gradient vanishing problem in deep networks, and helps to improve the recognition efficiency of the model. Experimental results on the NTU-RGBD120 dataset show that 3D-STARNET has significantly improved the accuracy of action recognition, and the top1 accuracy of the overall network reached 96.74%. This method not only solves the robustness shortcomings of existing methods, but also improves the ability to capture spatiotemporal features, providing an efficient and widely applicable solution for action recognition based on skeletal data.

Keywords: action recognition; spatiotemporal attention; multi-staged residual; skeleton; 3D CNN



Citation: Yang, J.; Sun, S.; Chen, J.; Xie, H.; Wang, Y.; Yang, Z. 3D-STARNET: Spatial–Temporal Attention Residual Network for Robust Action Recognition. *Appl. Sci.* **2024**, *14*, 7154. <https://doi.org/10.3390/app14167154>

Academic Editor: Douglas O’Shaughnessy

Received: 17 July 2024

Revised: 7 August 2024

Accepted: 12 August 2024

Published: 15 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the development of deep learning technology, action recognition has become an important research topic in the field of video understanding. It has extremely wide application scenarios in the fields of autonomous driving, safety control, rail transportation, human–computer interaction, etc. Currently, most behavior recognition methods are based on RGB [1], skeleton [2], optical flow [3], depth map [4], radar [5], point cloud, and other modal data or the fusion of these modalities. Skeleton data are usually robust to lighting changes, background interference and occlusion, and their computational complexity is much smaller than that of other modal data. Skeleton sequences also have two significant advantages in the field of behavior recognition [6]: (1) The spatial information within the frame contains the correlation between different joint points, and rich structural information can be extracted. (2) The temporal information between frames preserves rich temporal correlations, so this article conducts research based on skeletal data.

There are still many challenges in the field of action recognition using skeleton data. These challenges include the differences in the way different individuals perform the same

action, the complexity of complex backgrounds, and the fluctuation of lighting conditions. One of the key issues is to effectively mine the spatiotemporal features inherent in skeleton sequences. Currently, in the field of skeleton-based action recognition, the most popular method is based on graph convolutional networks (GCNs) [7]. It is built based on a series of skeleton graph sequences, where each node corresponds to a joint of the human body and defines two types of edges: spatial edges, which conform to the natural connectivity of joints, and temporal edges, which connect the same joints in consecutive time steps and allow information to be integrated along the spatial and temporal dimensions to discover action patterns in spatial and temporal dimensions. This not only improves the expression ability of the model, but also facilitates generalization in different contexts. Although GCN can explore action patterns in the spatiotemporal dimension, its recognition ability is significantly affected by the coordinate distribution, so it is not robust.

To address the above problems, we proposed the Spatial–Temporal Attention Residual Network for 3D human action recognition (3D-STARNET). In order to make the model robust, we convert the skeleton points into heat maps through Gaussian transformation, thus avoiding the strong dependence on the skeleton points. In order to enhance the network’s ability to extract key frames and key regions between frames, we propose a spatiotemporal attention mechanism (STA). In order to improve the efficiency of signal propagation in the network, we introduced a phased residual structure (MS-Residual). Finally, we verified 3D-STARNET on the NTU-RGBD120 dataset (see Section 4 for details), achieving significant improvements in accuracy.

To address the above issues, we proposed a Spatial–Temporal Attention Residual Network (3D-STARNET) for 3D human action recognition. Unlike GCN-based methods, we transform skeleton points into heat maps instead of skeleton sequences through Gaussian transformation. This change avoids a strong dependence on skeleton points, and this improvement significantly enhances the robustness of the algorithm. In order to enhance the network’s ability to extract key frames and key areas between frames, we proposed a spatiotemporal attention mechanism (STA). In order to improve the efficiency of signal propagation in the network, we introduced a staged residual structure (MS-Residual). Finally, we verified 3D-STARNET on the NTU-RGBD120 dataset (see Section 4 for details), and the accuracy was significantly improved.

2. Related Work

2.1. GCN-Based Action Recognition

Due to the significant advantages of the graph convolutional network (GCN) in processing graph data, it is widely used in the field of skeleton-based behavior recognition. It converts human skeleton sequences into spatiotemporal graphs and can simulate the complex spatiotemporal structure of human joints and correlation; therefore, GCN-based behavior recognition has become a popular research area [8]. Yan et al. [9] proposed a new dynamic skeleton model ST-GCN that can automatically learn temporal and spatial patterns from data. However, it ignores the implicit correlation between skeletons. Therefore, Li et al. [10] proposed the action-structure graph convolutional network (AS-GCN), which stacks graph convolution and temporal convolution as basic building blocks to learn the spatial and temporal features of behavioral actions. While performing action recognition, it can also predict future actions to help capture more detailed action patterns through self-supervision. In order to preserve the loss implicit joint correlations, Peng et al. [11] proposed an automatically designed GCN, explored the spatiotemporal correlations between nodes, and constructed a search space with multiple dynamic graph modules. They also proposed corresponding sampling and memory-efficient evolution strategies to search the space. The resulting architecture verified the effectiveness of high-order approximation and layer-by-layer dynamic graph modules. The G3D module proposed by Liu et al. [12] utilizes dense cross-spacetime edges as skip connections to directly propagate information in the space-time graph, and proposes a multi-scale aggregation scheme to reveal the importance of nodes in different neighborhoods for effective long-range modeling.

Tu et al. [13] designed a novel correlation-driven joint-skeleton fusion graph convolutional network (CD-JBF-GCN) as an encoder and used a pose prediction head as a decoder to achieve semi-supervised learning. The motion transfer between the joint stream and the bone stream can be explored, thereby promoting the two streams to learn more discriminative feature representations. In order to improve the flexibility of GCN in modeling temporal information, Liu et al. [14] proposed a temporal decoupled graph convolutional network (TD-GCN), which first extracts high-level spatiotemporal features from skeleton data, and then calculates the channel-dependent and time-dependent adjacency matrices corresponding to different channels and frames to capture the spatiotemporal dependencies between skeleton joints. Finally, in order to fuse the topological information of adjacent skeleton joints, the spatiotemporal features of skeleton joints are fused based on the channel-dependent and time-dependent adjacency matrices. Wang et al. [15] proposed a dynamic dense graph convolutional network (DD-GCN) that uses 4D adjacency modeling to construct a dense graph as a comprehensive representation of motion sequences at different levels of abstraction. Although a large number of studies have achieved promising results, GCNs are sensitive to noise or irregular connections between skeleton key points, which may affect the recognition accuracy, especially in complex actions.

2.2. Three-Dimensional CNN-Based Action Recognition

It is widely acknowledged that CNN has made great achievements in processing two-dimensional images, such as object detection [16,17] and instance segmentation [18]. However, it still faces considerable challenges in tasks based on skeletons and other objects with spatiotemporal information. In the field of skeleton-based action recognition, GCN has always occupied a mainstream position. In order to explore the potential of 3D CNN in capturing the spatiotemporal features of skeleton sequences, many researchers have conducted research on action recognition based on 3D CNN.

Ji et al. [19] pioneered a new 3D CNN model for action recognition, which extracts features from the temporal and spatial dimensions through 3D convolution, thereby capturing motion encoding information in multiple adjacent frames. However, Ref. [19] ignored the long-term spatiotemporal dependencies of videos. To solve this problem, Diba et al. [20] designed a temporal 3D convolutional network that can densely and efficiently capture short-, medium-, and long-term appearance and temporal information. Feichtenhofer et al. [21] introduced the SlowFast network for video recognition, in which the slow path of the network runs at a low frame rate to capture spatial semantics, and the fast path runs at a high frame rate to capture motion with fine temporal resolution. By reducing the channel capacity, the fast path can be made very lightweight and can fully learn temporal information. In their subsequent work X3D [22], a tiny 2D image classification architecture is gradually expanded in spatial, temporal, width, and depth dimensions. A simple step-wise network expansion method is used, which expands one axis at each step, achieving a good accuracy and complexity trade-off. In order to find a balance between model size and inference efficiency, Yang et al. [23] proposed a skeleton-based double-feature double-motion network (DD-NET), which achieves ultra-fast inference speed by lightweighting the network.

In order to construct 3D input data, previous studies either converted the pose distance matrices at different time spans into pseudo images and then stacked these pseudo images in time series [24], or summed up the 3D skeletons to obtain 3D data [25]. However, these methods suffer from the problem of information loss. Our study converts the skeleton sequence into a 2D heat map and then stacks it along the time dimension to form a 3D heat map volume to preserve all the information.

3. Proposed Methods

3.1. Model Overview

In this section, we introduce a skeleton-based 3D CNN human action recognition model called the Spatial–Temporal Attention Residual Network for 3D human action

recognition (3D-STARNet). Figure 1 shows the overall framework of the model. A spatiotemporal attention mechanism is embedded in the model and a staged residual structure is introduced. The spatiotemporal attention mechanism enables the network to enhance its feature learning ability in both temporal and spatial dimensions, while the staged residual structure improves the network's information transmission efficiency without changing the model's complexity. The overall network has been significantly improved in terms of robustness, scalability, and recognition performance. The details of 3D-STARNet will be introduced in the following chapters.

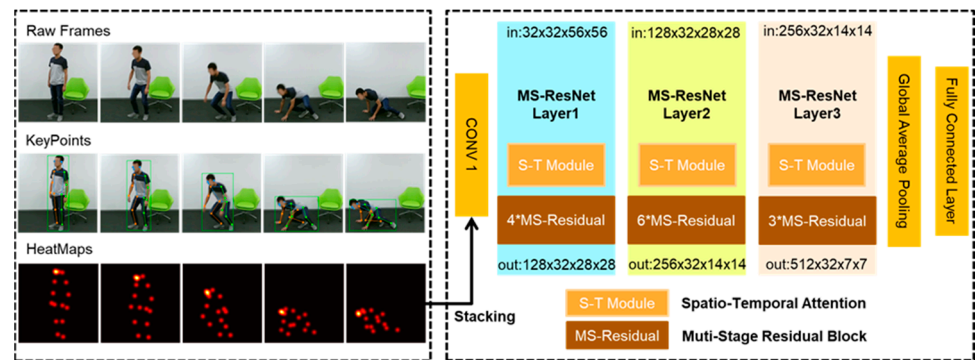


Figure 1. Visualization of 3D-STARNet input data processing and overall network architecture.

First, the conversion process of 2D skeleton key points to 3D heat maps is introduced, and then the two improvements of the article are introduced: the spatiotemporal attention mechanism and the multi-staged residual structure.

3.2. Skeleton Extraction and 3D Heat Map Generation

As the basic work of this study, the effect of skeleton key point extraction directly affects the accuracy of action recognition. This study first uses the mmpose [26] open source framework of openmmlab to perform top-down skeleton key point extraction, which is used as the basic condition for downstream tasks.

Since the basic action recognition framework used by the model in this article is pose3d [27], the data input format of its backbone network is a 3D heat map instead of a 2D skeleton sequence. This is mainly because a 3D heat map (the 3D heat map is the 2D heat map stacked in the time dimension so that it can be input into the 3D CNN model) has more time dimension information than a 2D skeleton sequence, which can better describe the temporal transformation relationship between actions. It has good robustness when dealing with occlusion, lighting changes, and complex backgrounds, can better handle noise and uncertainty in posture estimation, and reduces the dependence on precise coordinates of bones. At the same time, 3D heat maps can provide the model with more contextual information, which helps the model generalize between different scenes and different actions.

In this framework, we first use the pose estimator mmpose [26] to extract 2D human skeleton key points from action video clips, and then a Gaussian distribution heat map is generated for each skeleton key point to form a $K \times H \times W$ heat map, where K represents the number of skeleton points, and H and W are the height and width of the frame, respectively. These 2D heat maps are then stacked along the time dimension to form a 3D heat map volume of shape $K \times T \times H \times W$. Assuming that the position information of each bone point is represented by a triple (x_k, y_k, c_k) , the k joint points are mapped to a graph through Gaussian transformation to form a three-dimensional bone point heat map J :

$$J_{kij} = e^{-\frac{(i-x_k)^2 + (j-y_k)^2}{2\sigma^2}} * c_k, \quad (1)$$

Among them, x_k and y_k represent the position information of the k -th skeleton point, respectively, and c_k represents the confidence score of the skeleton point; this value is

obtained during the key point extraction stage. Its value is usually between 0 and 1, representing the accuracy of detecting the skeleton point. It affects the size of the variance. The larger the variance, the wider the Gaussian expansion range, which will be more advantageous in tasks with strong globality such as action recognition. And (i, j) represents the point on the Gaussian graph, where σ represents the degree of diffusion around the key point, ensuring that the generated Gaussian kernel can cover the area around the key point, which will be more advantageous in tasks with strong globality such as action recognition. Figure 2 intuitively shows the transformation from an RGB image to skeleton points and then to a heat map.

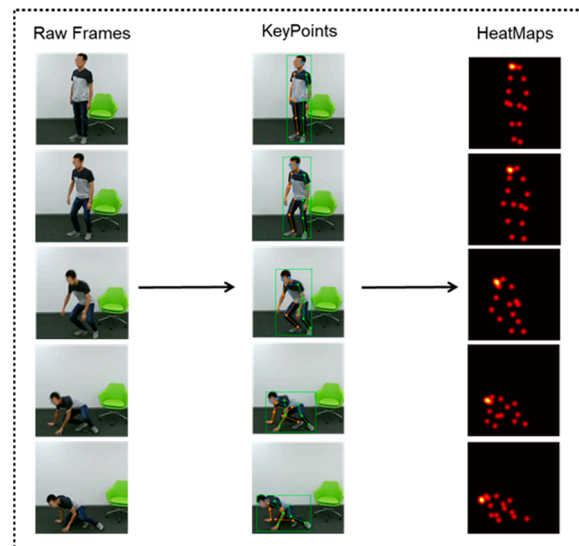


Figure 2. Data acquisition process.

3.3. Spatiotemporal Attention Module

The attention mechanism is a cognitive enhancement process that enables the network to selectively learn interesting feature areas while ignoring other less relevant features. The input data of this paper are a continuous action video frame. The ordinary network assigns the same weight to all frames and pixel positions within the frame, which brings great confusion to the action recognition task. To this end, this paper proposes a spatiotemporal attention mechanism STA, as shown in Figure 3. Our proposed spatiotemporal attention mechanism applies attention mechanisms in both temporal and spatial dimensions to enhance the model's ability to identify key frames and key locations in specific frames among consecutive video frames.

F denotes the input continuous video frames, where each frame is represented as a feature map $f_{i,j}$ in $R^{h \times w}$, with $i = 1, 2, \dots, l$ indicating the time index, and $j = 1, 2, \dots, c$ indicating the channel index. Here, l is the total number of video frames and c is the number of channels. Through the Spatiotemporal Attention (STA) module, F can generate corresponding weights $W = \{w_{i,j}\}$ in both the temporal and spatial dimensions. We define the attention function T as a mechanism to learn these weights W from F , which in turn allows us to define the output feature $O = \{o_{i,j}\}$ in $R^{h \times w}$ after F has passed through the STA module. It should also be noted that T is composed of two parts, T_t and T_s , where T_t and T_s represent the temporal attention function and the spatial attention function, respectively. Figure 4 shows the entire spatiotemporal attention mechanism. The construction of the entire attention mechanism follows the principles of making the architecture simple and effective enough and making the network robust.

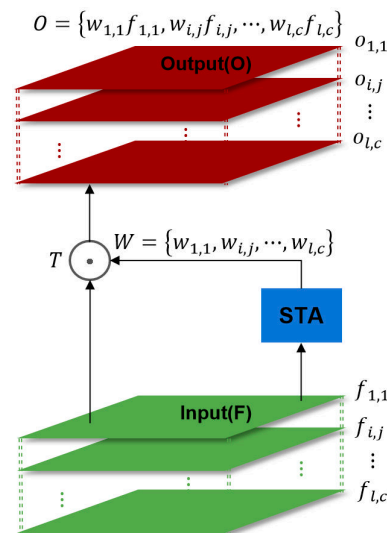


Figure 3. STA module architecture embedded in 3D CNN model.

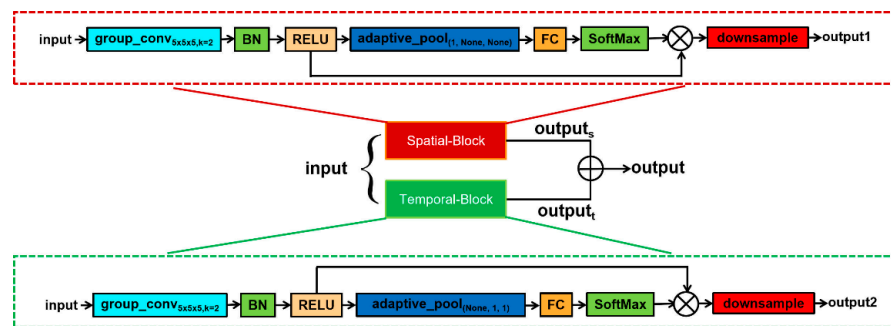


Figure 4. STA module architecture.

By introducing the spatiotemporal attention mechanism we proposed, the network can recognize actions in the video by focusing on only a small part of the frames in the video. This is difficult to do with previous 3D CNN models. Previous studies mostly focused on spatial and channel dimensions, while giving indifferent weights to video frames, which made it difficult for the network to have stable recognition performance and greatly increased the processing cost of the network. The temporal attention mechanism in the STA module can effectively enhance the model's ability to identify key frames, thereby more effectively extracting decisive information about video actions. Specifically, the input feature F_t learns the weight of the time dimension through the attention function T_t . In practical applications, since the input frame sequence is too short, it is impossible to extract the variance information of the frame sequence. To address this problem, we use a group convolution to expand the time channel to twice the original size to retain richer time information. Here, the attention function T_t can be defined as a composite function that generates temporal weights for each frame:

$$T_t = S_t \circ \varphi_t \circ \varepsilon_t \quad (2)$$

where S_t and ε_t represent the compression and expansion operations of the time channel, respectively, φ_t represents the Relu activation function, and \circ represents the composite operation of multiple functions; that is, the output of one function becomes the input of the next function.

The upper part of Figure 4 shows the detailed network framework of the temporal attention module. The input data first undergo group convolution with a grouping of 2 to expand the temporal dimension, and then undergo BN and Relu to achieve batch normalization of the data and introduce nonlinear features. Then, a pooling operation is

performed on the temporal dimension, and the temporal weight is obtained through a fully connected layer and a SoftMax activation function, and finally a weighted operation is performed with the original input data.

The 3DCNN model also includes information about the spatial dimension. Similar to the temporal attention mechanism, the spatial attention mechanism is used to obtain the weights of the spatial dimension, enabling the model to focus on a certain area on a single frame, which is also very helpful for accurately identifying behavioral actions. Specifically, for the input feature F_s , the spatial attention mechanism learns the weights corresponding to each channel in the spatial dimension through the attention function T_s . The spatial attention mechanism also includes an expansion and compression operation of the spatial dimension. The attention function T_s can also be defined as a composite function that generates a weight for each spatial channel of each frame:

$$T_s = S_s \circ \varphi_s \circ \varepsilon_s \quad (3)$$

where S_s and ε_s represent the compression and expansion operations of the spatial channel, respectively, and φ_s represents the Relu activation function.

The lower part of Figure 4 shows the detailed network structure of the spatial attention mechanism. It can be easily seen that it is very similar to the temporal attention mechanism, except that the pooling dimension in the adaptive pooling layer is different.

Finally, in order to enable the model to obtain key frames in continuous video frames and learn key areas in a single frame, we combined temporal attention with spatial attention to form a spatiotemporal attention module STA, as shown in Figure 4. Assuming that output_s and output_t represent the output of spatial attention and temporal attention, respectively, and the output represents the combination of the above two parts to represent the output of STA. In STA, after mixed compression operations on the temporal dimension and the spatial dimension, the ability to obtain information from a global scope is realized.

3.4. MS-Residual Structure

In order to make the model competitive in the efficient dissemination of information, we improved the previous Resnet [28] residual block into a phased residual block, and divided it into start block, middle block, and end block, without changing the complexity of the model. Here, we only changed the order of the different units to promote the dissemination of data in the network. At the same time, since a shortcut exists in the main information dissemination path, we have also conducted research on its improvement. The following is a detailed introduction to the improvement.

The residual structure of all layers of the traditional resnet network architecture adopts the same unit structure, as shown in Figure 5. The propagation of information in the residual block can be expressed as follows:

$$X_{i+1} = \begin{cases} X_i + F(X_i, W_i), & \text{size}(F(X_i, W_i)) = \text{size}(X_i) \\ \lambda_i X_i + F(X_i, W_i), & \text{size}(F(X_i, W_i)) \neq \text{size}(X_i) \end{cases} \quad (4)$$

where X_i and X_{i+1} represent the input and output information of the i -th residual block, respectively, and the function F represents a learnable residual mapping function, which represents the downsampling operation performed when the sizes of $F(X_i, W_i)$ and X_i are different.

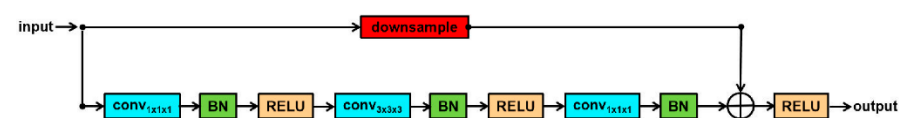


Figure 5. The original residual structure.

As can be seen from Figure 5, the outputs of the main branch and shortcut are added and then processed using RELU which will cause the negative elements in the information

to be zero, thereby causing information loss. Pre-activation [29] has improved this by moving the BN and RELU at the end of the network to the front of the network to obtain better performance. However, this improvement does not achieve a normalization of the complete signal, but is only applied to the branch signal, which reduces the information control through the network and affects the dynamic flow of the gradient. As the number of network layers increases, the overall information lacks normalization and nonlinearity.

To address the above issues, we replace the original residual structure with a staged residual structure [30], as shown in Figure 6. The same number and type of unit blocks as the original residual structure, but the order of the unit blocks is changed. Specifically, in the start stage, the last Relu operation is removed to prevent the loss of signal caused by the zeroing of negative numbers. In the middle stage, we adopted the residual structure in pre-activation. In the end stage, we said that BN and Relu are applied to the complete signal to replace the last BN and Relu of each residual block in each stage, so as to effectively control the complete signal. When the gradient of the signal becomes very large after passing through the residual block in the initial stage, it will also be subjected to nonlinear normalization processing using BN and Relu in the final stage to improve the nonlinear expression ability of the model and reduce the overall signal loss.

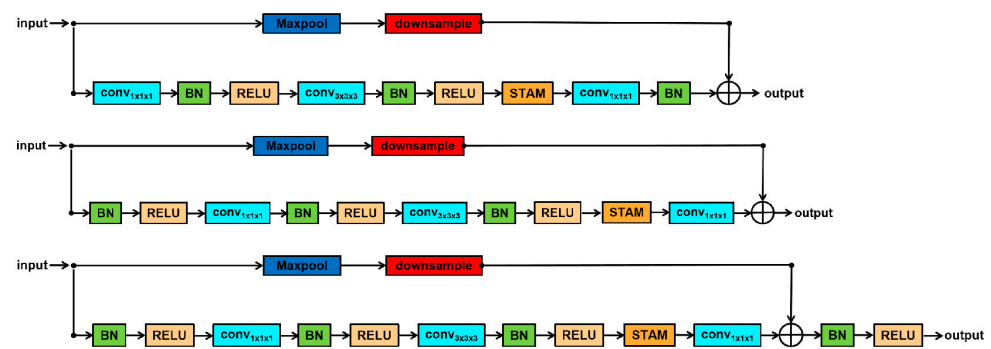


Figure 6. Multi-stage residual structure.

In the residual structure, when the dimension of the input X_i is inconsistent with the mainstream output dimension, a shortcut projection method is used to make the second condition in Formula (4) meet. The original ResNet's shortcut projection is implemented by a convolution with a kernel of $1 \times 1 \times 1$ and a stride of 2, which is used to satisfy the condition of adding operations with the mainstream branch. For an input dimension of $T \times H \times W$, after the above convolution operation, the output dimension will be changed to $\frac{T}{2} \times \frac{H}{2} \times \frac{W}{2}$. Therefore, this convolution operation will cause about 87.5% of the information to be lost, and the remaining 12.5% of the information will be meaningless, which is equivalent to adding noise to the mainstream output and will have a negative impact on network performance.

To address the above problems, we have improved the shortcut projection method, as shown in Figure 7. In this improvement, we separated the spatial projection and channel projection from the projection, and used MaxPooling with a stride of $1 \times 2 \times 2$ and a kernel_size of $1 \times 3 \times 3$ to perform projection mapping in the spatial and temporal dimensions to emphasize the significant features in the local area, suppress unimportant features, and increase the translation invariance of the input data. Then, a $1 \times 1 \times 1$ convolution is applied to the channel dimension, and finally BN is applied to normalize the data. In addition to reducing information loss, this improvement can also effectively improve the translation invariance of the network.

The improvements we introduced above, including the improvement of residual structure and shortcut, are all carried out without increasing the model parameters and complexity, and the performance of the final network is also significantly improved.

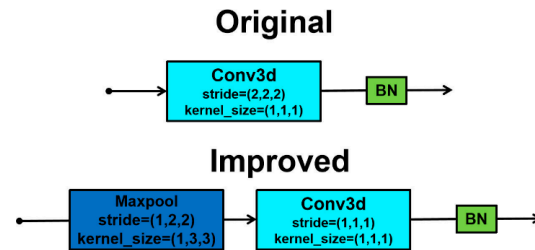


Figure 7. Comparison before and after shortcut improvement.

4. Experiments

We comprehensively evaluate our proposed 3D-STARNET architecture on NTU-RGBD120 and analyze the performance of each improvement.

4.1. Datasets and Evaluation Metrics

The NTU-RGBD120 [31] dataset is a large-scale 3D human action recognition dataset developed to address some limitations in existing action recognition benchmarks, including the lack of large-scale training samples, the limited number of different categories of actions, the insufficient diversity of camera viewpoints, the limited changes in environmental conditions, and the insufficient diversity of human subjects. It also provides multiple data modalities, including RGB, skeleton points, depth map, and infrared.

In this study, we only used the skeleton modality. In the action category, we selected 12 actions, including throwing, kicking something, squatting down, moving heavy objects, cross arms, falling, punching other person, kicking other person, pushing other person, wielding knife towards other person, hitting with body, grabbing other person's stuff. The action description is shown in Table 1. In addition, in order to verify the generalization ability of the model, we also used another dataset—UCF101 [32]. UCF101 is a real action dataset that provides 13,320 videos from 101 action categories. We used acc/top1 as the accuracy indicator of the model, and used the confusion matrix to intuitively understand the recognition effect of the model.

The improvements we introduced above, including the improvement of residual structure and shortcut, are all carried out without increasing the model parameters and complexity, and the performance of the final network is also significantly improved.

Table 1. Action number and description.

Number	Action Description	Number	Action Description
A	throwing	G	punching other person
B	kicking something	H	kicking other person
C	squatting down	I	pushing other person
D	moving heavy objects	J	wielding knife
E	cross arms	K	hitting with body
F	falling	L	grabbing stuff

4.2. Implementation Details

We used Python to develop programs and build models based on the pytorch framework, running them on a server with an RTX 4090 GPU and training them on the behavior recognition framework MMAAction2. For the 3D CNN model, our model is divided into three stages, with four, six, and three basic network blocks, respectively. The basic network block is the bottleneck block of resnet. For the proposed spatiotemporal attention, we set each network block to achieve a balance between performance and model size. For the second improvement point, since the size of the model does not change, we made improvements at each stage. For input data, we used RepeatDataset as a wrapper to repeat the dataset and reshape the 3D heat map to $48 \times 56 \times 56$. In this study, we used the stochastic gradient descent (SGD) optimizer to train the deep learning model, and the total number

of training rounds was set to 30 rounds in order to achieve fast and stable convergence. The configuration of the optimizer includes the learning rate set to 0.01, the momentum parameter set to 0.9, and the weight decay parameter set to 0.0003, which is a regularization method to help prevent the model from overfitting. In addition, in order to effectively control the gradient explosion problem that may occur during training, we introduced the gradient clipping technique, in which the maximum L2 norm of the gradient is limited to 40.

4.3. Experiments on the NTU-RGBD120 Dataset

We experimented with the 3D-STARNET model on the NTU-RGBD120 and UCF101 dataset, using acc/top1 as the model's accuracy metric, and compared it with other advanced CNN-based and GCN-based behavior recognition models. As shown in Table 2, our model is ahead of them in terms of accuracy.

Table 2. Comparison of different methods on NTU-RGBD120 and UCF101 from acc/top1.

Model	Acc/Top1 (NTU-RGBD120)	Acc/Top1 (UCF101)
posec3d [27]	93.03%	93.10%
x3d [22]	94.07%	91.13%
2s-agcn [2]	92.97%	85.22%
Stgcn [8]	93.62%	90.79%
3D-STARNET	96.74%	93.96%

Figure 8 uses a confusion matrix to intuitively reflect the detection results of our model on the 12 types of actions in the validation set. A to L in this figure represent the 12 categories mentioned above; please refer to Table 1 for details.

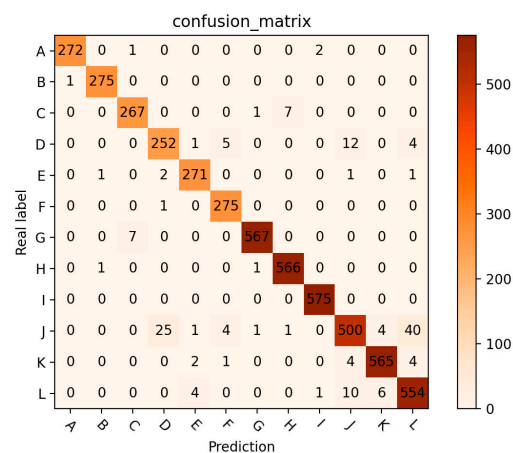


Figure 8. Confusion matrix.

4.4. Ablation Experiments

In this section, we conducted sufficient ablation experiments to verify the effectiveness of various improvements in 3D-STARNET. The results of the ablation experiments are shown in Table 3.

Table 3. Ablation experiment results.

Model	Acc/Top1 (NTU-RGBD120)	Acc/Top1 (UCF101)
Posec3d	93.03%	93.10%
Posec3d + ST-Module	96.15% + 3.12%	93.12% + 0.02%
Posec3d + MS-Residual	95.51% + 2.48%	93.60% + 0.5%
3D-STARNET	96.74% + 3.71%	93.96% + 0.86%

First, the original ResNet Bottleneck backbone network was improved by changing it to a staged residual structure. With the number of parameters unchanged, the accuracy was improved by 2.48% on the NTU-RGBD120 dataset and by 0.5% on the UCF01 dataset. Another experiment introduced the spatiotemporal attention module into the ResNet network model, which achieved a 3.12% accuracy improvement on the NTU-RGBD120 dataset and kept the accuracy basically the same on the NCF101 dataset. Finally, the above two improvements were introduced on the basis of ResNet, which improved the overall network in terms of information transmission efficiency and model feature focusing ability, achieving 3.71% and 0.86% accuracy improvements in the two datasets, respectively.

5. Conclusions

The focus of this study is to use skeleton data to identify human behavior. We introduced the skeleton-based action recognition model 3D-STARNET to achieve a robust detection of action. Firstly, 3D heat maps were used instead of 2D skeleton sequences as the input of the model to reduce the strong dependence on skeleton points. Secondly, we proposed a spatiotemporal attention mechanism to improve the model's information processing and feature extraction capabilities in the spatiotemporal dimension. Finally, we introduced a multi-stage residual structure to achieve efficient information transmission without changing the computational complexity. We verified it on the NTU-RGBD120 dataset, and the overall accuracy reached 96.74%, which has broad application prospects in monitoring and early warning, safety prevention and control, rail transit, and human–computer interaction.

Although 3D-STARNET has achieved remarkable results in action recognition, we hope that the model has generalization capabilities in action recognition in different environments and different complexities. Future research can further explore the performance of the model in diverse scenarios and explore the fusion of skeletal data with other types of data, such as RGB images or depth information, to further improve the accuracy and robustness of behavior recognition.

Author Contributions: The authors confirm contribution to the paper as follows: methodology, J.Y. and S.S.; writing—original draft preparation, J.Y. and S.S.; resources, J.Y.; writing—review and editing, J.Y., S.S., J.C. and Z.Y.; data curation, J.C.; software, H.X. validation, Z.Y.; investigation, S.S., J.C., H.X., Y.W. and Z.Y.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the National Special Project of Science and Technology Basic Resources Survey (grant No. 2022FY101400) and the National Natural Science Foundation of China Innovation Group Project (grant No. 52121003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The NTU-RGBD 120 dataset is available at <https://rose1.ntu.edu.sg/dataset/actionRecognition/> (accessed on 24 June 2024).

Acknowledgments: We would sincerely like to thank the people who supported this work and the reviewing committee for their estimable feedback.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

1. Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; Liu, J. Human action recognition from various data modalities: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 3200–3225. [[CrossRef](#)] [[PubMed](#)]
2. 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.

3. Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M.J. On the integration of optical flow and action recognition. In Proceedings of the Pattern Recognition: 40th German Conference, GCPR 2018, Stuttgart, Germany, 9–12 October 2018; pp. 281–297.
4. Baek, S.; Shi, Z.; Kawade, M.; Kim, T.-K. Kinematic-layout-aware random forests for depth-based action recognition. *arXiv* **2016**, arXiv:1607.06972.
5. Kim, Y.; Ling, H. Human activity classification based on micro-Doppler signatures using a support vector machine. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 1328–1337.
6. Ren, B.; Liu, M.; Ding, R.; Liu, H. A survey on 3d skeleton-based action recognition using learning method. *Cyborg Bionic Syst.* **2024**, *5*, 0100. [[CrossRef](#)] [[PubMed](#)]
7. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
8. Zhang, X.; Xu, C.; Tian, X.; Tao, D. Graph edge convolutional neural networks for skeleton-based action recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 3047–3060. [[CrossRef](#)] [[PubMed](#)]
9. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence, New Orleans, LA, USA, 2–7 February 2018.
10. 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.
11. Peng, W.; Hong, X.; Chen, H.; Zhao, G. Learning graph convolutional network for skeleton-based human action recognition by neural searching. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 2669–2676.
12. Liu, Z.; Zhang, H.; Chen, Z.; Wang, Z.; Ouyang, W. Disentangling and unifying graph convolutions for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 143–152.
13. Tu, Z.; Zhang, J.; Li, H.; Chen, Y.; Yuan, J. Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition. *IEEE Trans. Multimed.* **2022**, *25*, 1819–1831. [[CrossRef](#)]
14. Liu, J.; Wang, X.; Wang, C.; Gao, Y.; Liu, M. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE Trans. Multimed.* **2023**, *26*, 811–823. [[CrossRef](#)]
15. Wang, X.; Zhang, W.; Wang, C.; Gao, Y.; Liu, M. Dynamic dense graph convolutional network for skeleton-based human motion prediction. *IEEE Trans. Image Process.* **2023**, *33*, 1–15. [[CrossRef](#)] [[PubMed](#)]
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [[CrossRef](#)]
18. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 3523–3542. [[CrossRef](#)] [[PubMed](#)]
19. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [[CrossRef](#)] [[PubMed](#)]
20. Diba, A.; Fayyaz, M.; Sharma, V.; Karami, A.H.; Arzani, M.M.; Yousefzadeh, R.; Van Gool, L. Temporal 3d convnets: New architecture and transfer learning for video classification. *arXiv* **2017**, arXiv:1711.08200.
21. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6202–6211.
22. Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 203–213.
23. Yang, F.; Wu, Y.; Sakti, S.; Nakamura, S. Make skeleton-based action recognition model smaller, faster and better. In Proceedings of the 1st ACM International Conference on Multimedia in Asia, Beijing, China, 16–18 December 2019; pp. 1–6.
24. Lin, Z.; Zhang, W.; Deng, X.; Ma, C.; Wang, H. Image-based pose representation for action recognition and hand gesture recognition. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), Buenos Aires, Argentina, 16–20 November 2020; pp. 532–539.
25. Liu, H.; Tu, J.; Liu, M. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv* **2017**, arXiv:1705.08106.
26. Sengupta, A.; Jin, F.; Zhang, R.; Cao, S. mm-Pose: Real-time human skeletal posture estimation using mmWave radars and CNNs. *IEEE Sens. J.* **2020**, *20*, 10032–10044. [[CrossRef](#)]
27. Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; Dai, B. Revisiting skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2969–2978.
28. 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.
29. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.

30. Duta, I.C.; Liu, L.; Zhu, F.; Shao, L. Improved residual networks for image and video recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 9415–9422.
31. 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. [[CrossRef](#)] [[PubMed](#)]
32. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.

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.