applied sciences

MDPI

*Article*

# Multi-Head TrajectoryCNN: A New Multi-Task Framework for Action Prediction

**Xiaoli Liu and Jianqin Yin ***

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China; liuxiaoli134@bupt.edu.cn

* Correspondence: jqyin@bupt.edu.cn

**Abstract:** Action prediction is an important task in human activity analysis, which has many practical applications, such as human–robot interactions and autonomous driving. Action prediction often comprises two subtasks: action semantic prediction and future human motion prediction. Most of the existing works treat these subtasks separately, ignoring the correlations, leading to unsatisfying performance. By contrast, we jointly model these tasks and improve human motion predictions utilizing their action semantics. In terms of methodology, we propose a novel multi-task framework (Multi-head TrajectoryCNN) to simultaneously predict the action semantics and human motion of future human movements. Specifically, we first extract a general spatiotemporal representation of partial observations via two regression blocks. Then, we propose a regression head and a classification head for predicting future human motion and action semantics of human motion, respectively. For the regression head, another two stacked regression blocks and two convolutional layers are applied to predict future poses from the general representation learning. For the classification head, we propose a classification block and stack two regression blocks to predict action semantics from the general representation. In this way, the regression and classification heads are incorporated into a unified framework. During the backward propagation of the network, the human motion prediction and the semantic prediction may be enhanced by each other. NTU RGB+D is a widely used large-scale dataset for action recognition, which was collected by 40 different subjects from three views. Based on the official protocols, we use the skeletal modality and process action sequences with fixed lengths for the evaluation of our action prediction task. Experiments on NTU RGB+D show our model's state-of-the-art performance. Furthermore, the experimental results also show that semantic information is of great help in predicting future human motion.

**Keywords:** action prediction; human motion prediction; multi-task framework

check for updates

**Citation:** Liu, X.; Yin, J. Multi-Head TrajectoryCNN: A New Multi-Task Framework for Action Prediction. *Appl. Sci.* **2022**, *12*, 5381. https://doi.org/10.3390/app12115381

Academic Editor: Cezary Czaplewski

Received: 20 April 2022
Accepted: 24 May 2022
Published: 26 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Copyright:** © 2022 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

Predicting future action information is important due to its undeniable applications in computer vision and robotics, among which action semantics and human motion predictions are two key aspects [1]. As shown in Figure 1, our prediction task is different from the traditional recognition task. Our prediction task aims to predict action-related information before the action completes its execution, including action semantics and future human motion.

Due to the limited action information of partial observed sequences, the action prediction task is particularly challenging. As shown in Figure 1, the action execution of the partial observed sequence is incomplete. Its semantics could be ambiguous, such as "Directing" and "Basketball Shots", and its future movement trend may be uncertain. Aiming at these challenges, these works [2,3] predict multiple plausible human motion sequences. Unlike these tasks, we aim at predicting future action information accurately using a limited partial observed sequence as soon as possible. Therefore, to address this challenge, the action semantics and human motion predictions can be associated with each

other. For example, the action semantic information can reduce future motion uncertainty, which helps guide the model to predict a more natural motion sequence. Conversely, future motion information can help reduce the ambiguity of the action semantics.

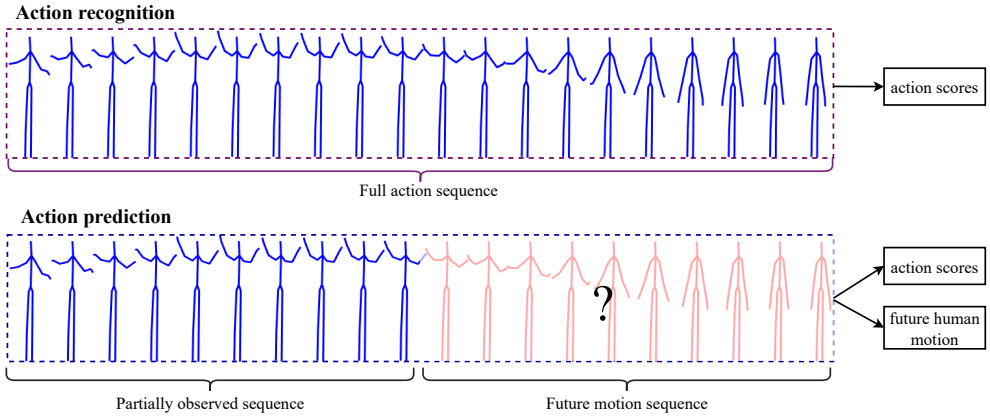



**Figure 1.** Human activity analysis.

Most of the existing works model action semantic prediction and future human motion prediction separately, which ignores the relationships between action semantics and human motion, thus leading to unsatisfied predictions. For example, these works [4,5] focus on predicting a short certain future human motion sequence using on a short historical human motion sequence. Furthermore, these works [6,7] focus on predicting the action semantics of an incomplete human motion sequence very early. Due to the action ambiguous semantics and uncertainty of future motion caused by limited observed information, it is challenging to separately predict action semantics and certain future motion sequences, leading to poor performance of these models. By contrast, we jointly model action semantics and future human motion predictions. In this way, these tasks can be associated with each other in a unified framework, which helps to reduce the semantic ambiguity and uncertainty of future motion.

In this paper, we propose a new multi-task framework, Multi-head TrajectoryCNN, to jointly model action semantic and human motion prediction tasks. Specifically, we first introduce a trajectory block proposed in [8] as our primary building block, named the regression block. Moreover, we build an encoder using stacked regression blocks to extract general spatiotemporal features of the input sequence. Furthermore, the extracted general spatiotemporal features are shared by the classification and regression heads for computational efficiency. Finally, we build a regression head and a classification head for predictions of future human motion and its action semantics. For the regression head, we first build a decoder with the regression block to predict future poses. For the classification head, we build another decoder using a modified trajectory block and fully connected (FC) layers to predict the action semantics of human motion. Instead of using the original trajectory block, we propose a modified trajectory block named a classification block for the classification head, which shows the effectiveness of the proposed modified trajectory block. With these elegant designs, the regression head and the classification head are connected together, using each other's information to reduce the uncertainty and semantic ambiguity caused by the partial observed motion sequence.

Our main contributions can be summarized as follows.

(1) We improve the model performance of human motion prediction utilizing action semantics, which helps reduce the uncertainty of future movements.
(2) We propose a new multi-task framework (i.e., Multi-head TrajectoryCNN) to jointly predict the action semantics and human motion of future motion dynamics. Different tasks of action prediction are incorporated in a unified model, which associates these tasks to improve their performance with each other.

(3) Experiments on the NTU RGB+D dataset show the state-of-art performance, proving the effectiveness of our proposed method. The experimental results also show that the action semantics can greatly help predict accurate future poses, and the average errors of our method reduce by 10.12 mm and 11.07 mm per joint for CV and CS protocols.

## 2. Related Work

We investigated the related works from these aspects: human motion prediction, action semantic prediction, and multi-task frameworks for human activity analysis.

### 2.1. Human Motion Prediction

Three main types of methods are proposed to predict future human motion, including recurrent neural network (RNN)-based methods, convolutional neural network (CNN)-based methods, and graph convolutional network (GCN)-based methods.

For RNN-based methods, human poses are set as states of RNN cells, and the dynamics of pose evolution are modeled via state–state transitions of the cells. Fragkiadaki et al. [9] proposed an encoder-recurrent-decoder (ERD) model by incorporating a nonlinear encoder and decoder before and after long short-term memory (LSTM) cells. Martinez and Gui et al. [10,11] proposed gated recurrent unit (GRU)-based encoder–decoder models for human motion predictions. Chiu et al. [12] proposed a triangular-prism RNN (TP-RNN) for predictions, which models multi-scale features of human motion using a hierarchical framework. However, traditional RNN ignores the spatial modeling, usually resulting in unsatisfying performance. Some recent works incorporated graph representation and quaternion representation for the human body to enhance the spatial modeling of RNN models [13,14].

For CNN-based methods, the human motion sequences are usually represented as a 3D tensor. Then, CNN is applied to scan the width, height, and channel of the 3D tensor for modeling the spatiotemporal dynamics of human motion sequences [15,16]. For example, Li et al. [17] first represented the human motion sequence as a 3D tensor with the width as joints, the height as time, and the channel as the joint dimension. Then, the authors proposed a convolutional sequence-to-sequence model by taking the 3D tensor as the input, where the convolutional kernels scan the 3D tensor to model the local spatiotemporal features. With the hierarchical representation of multiple convolutional layers, spatiotemporal information with a larger receptive field was extracted for predictions. Liu et al. [8] also represented the motion sequence as a 3D tensor, but unlike [17], the channel is set as time. In this way, the global trajectory information is easily modeled by their CNN model, leading to its state-of-the-art performance.

For GCN-based methods, the human pose is often represented as a graph with joints as nodes and joint spatial relationship as edges [18]. In these models, the spatial information is modeled by updating the node and edge information of the graph, and the temporal information is usually modeled by discrete cosine transform (DCT). Mao et al. [19] represented the human pose as a fully connected graph where the edges connected every pair of joints. Moreover, the authors transformed the human motion sequence in the trajectory space using DCT and stacked several graph convolutional layers to extract the spatiotemporal features of human motion in the trajectory space. Cui et al. [20] also designed a GCN-based encoder and decoder for predicting human motion. In contrast to [19], the authors dynamically learned inherent and meaningful links of natural connectivity of the human body.

### 2.2. Action Semantic Prediction

Reviewing related works, three pipelines are adopted for action semantic prediction, including recognition-based action semantic prediction, future-representation-based action semantic prediction, and knowledge-distillation-based action semantic predictions.

Recognition-based methods mainly adopt the pipeline of traditional action recognition, which struggles to directly predict action semantics from the partial observed sequence [21,22].

For example, Aliakbarian et al. [23] developed a multi-stage LSTM model to leverage the context-aware and action-aware features for predicting action-semantic labels very early. Chen et al. [24] proposed a deep reinforcement learning (PA-DRL) model to activate the action-related parts for more accurate semantic prediction. Wu et al. [25] proposed a spatial-temporal relation reasoning approach for predictions, which utilizes a gated graph neural network to model the spatial relationships of visual objects and uses a long short-term graph network to model the short-term and long-term dynamics of the spatial relationships. However, recognition-based methods ignore the characteristics of the problem, which inevitably leads to poor performance due to the semantic ambiguity of partial observations.

Future-representation-based methods obtain more actionable information to reduce the ambiguity of action semantics by predicting future action information from partial observations [26,27], which is one of the methods to address the limitations of recognition-based methods. For example, Gammulle et al. [28] proposed a jointly learned model to jointly predict future visual spatiotemporal representation and its action semantics. Pang et al. [29] proposed a bidirectional dynamics network to jointly synthesize future motion and reconstruct the action information of partial videos. Similarly, Fernando et al. [30] proposed new measures to reconstruct the representation of partial videos by correlating the past with the future. In conclusion, future representation methods obtain improved performance utilizing future action information mined from partial observations, which helps to reduce the action ambiguity of partial videos caused by limited observations.

Knowledge-distillation-based methods usually distill completed action information from full videos to partial videos via the teacher–student framework [31], which is another method used to address the limitations of recognition-based methods. For example, Kong et al. [32] proposed a marginalized stacked autoencoder (MSDA)-based deep sequential context network (DeepSCN) to mine the rich sequential context information of partial videos and reconstruct completed action features from full videos. Wang et al. [7] proposed a progressive teacher–student learning framework to distill progressive action knowledge from full videos to partial videos. Similarly, Cai et al. [33] proposed a two-stage learning framework (knowledge transfer framework) to transfer the action knowledge from full videos to partial videos, which assists in achieving accurate predictions.

### *2.3. Multi-Task Framework for Human Activity Analysis*

Several recent works have focused on multi-task human activity analysis. Cai et al. [34] proposed a two-stage deep framework to jointly model the tasks of action generation, action prediction, and action completion in videos. Mahmud et al. [35] and Zhao et al. [36] jointly modeled future action semantic labels and their start time. Liao et al. [37] proposed a multi-task context-aware recurrent neural network (MCARNN) to predict future action and the location of the human. Unlike these multi-task frameworks, we focus on jointly predicting future human motions and their action semantics.

Recently, some similar works were proposed for jointly modeling future human motions and action semantics. For example, Guan et al. [38] proposed a deep generative model to learn a generative hybrid representation for action prediction, which jointly predicts future motions and action semantics. Liu et al. [39] jointly modeled hand motion and hand action in egocentric videos. Li et al. [40] proposed a non-autoregressive model to predict future human motion and action semantics. Similarly, Li et al. [41] proposed a symbiotic graph neural network for human motion prediction and action semantic prediction tasks.

## 3. Methodology

Given a partially observed motion sequence $\mathbf{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_N\} \in R^{N \times J \times D}$, its corresponding future motion sequence can be represented as $\mathbf{X}_2 = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \cdots \mathbf{x}_{N+M}\} \in R^{M \times J \times D}$ and its action semantic denoted by $y$, where $N$ is the length of the partially observed motion sequence, $M$ is the length of the future human motion sequence, $J$ is the number of joints of the human body, $D$ is the dimension of the joints (i.e., $D$ = 3),

$\mathbf{x}_i = \{j_{i1}, j_{i2}, \cdots, j_{iJ}\} \in R^{J \times D} (i = 1, 2, \cdots, N + M)$, and $j_{i*}$ is the $*$-th joint of the $i$-th pose. The action prediction model aims to predict the future motion sequence and action semantics using on a partially observed motion sequence, which can be formulated by Equation (1).

$$\{\mathbf{X}_2, y\} = f(\mathbf{X}_1) \tag{1}$$

where $f(\cdot)$ denotes the action prediction model.

In this paper, we parameterize the action prediction model via a multi-task framework, including a regression head and a classification head. The regression head predicts the future dynamics $\mathbf{X}_2$, and the classification head aims to predict the semantic label $y$.

### 3.1. Backbone Layers

To build our model efficiently, we propose two backbone layers, as shown in Figure 2, including a regression block and a classification block. The regression block is directly adopted from [8], and the classification is improved based on the regression block. As denoted in Figure 2, the regression block and the classification block share the same structure, and the operations within the blocks are different. Details are described below.

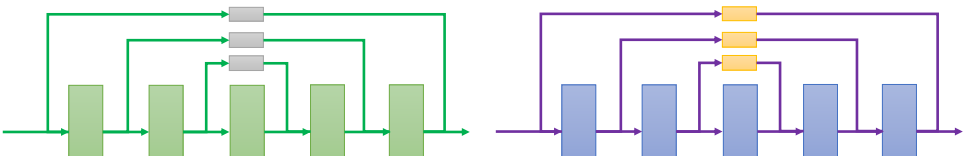


**Figure 2.** Backbone layers. The green boxes consist of a convolutional layer with a filter of $3 \times 3$, a Leaky ReLU layer, and a Dropout layer. The gray boxes consist of a convolutional layer with a filter of $1 \times 1$ and a Leaky ReLU layer. The blue boxes consist of a convolutional layer with a filter of $3 \times 3$, a ReLU layer, a MaxPooling layer, and a Dropout layer. The yellow boxes consist of a convolutional layer with a filter of $1 \times 1$, a ReLU layer, and multiple MaxPooling layers that match the dimension of the corresponding outputs of blue boxes.

Regression Block (RB). The architecture of the regression block is directly adopted from the trajectory block proposed in [8], which mainly consists of five convolutional layers. In the regression block, due to different sizes of filters, the convolutional layers in the main pipeline (i.e., the green boxes) model the joint-level features, and the convolutional layers in the residual connections (i.e., the gray boxes) model the point-level features (i.e., the axis-level features, which correspond to different axes of 3D joints). The convolutional layers in the main pipeline are connected between lower and deeper layers via residual connections. In this way, the regression block easily captures multi-scale features of human motion with an enlarged receptive field.

Classification Block (CB). Considering the different tasks, the classification block is designed for classification tasks that improve based on the regression block. Unlike the regression block, the MaxPooling layers are applied to the convolutional layers in the classification block, which aims to increase the receptive fields of the model to capture rich semantic features for early predictions of action semantics.

### 3.2. Network Structure

Based on the proposed backbone layers, we designed different encoders and decoders to build our Multi-head TrajectoryCNN, as shown in Figure 3. Firstly, an encoder is applied to extract a general spatiotemporal representation of the input motion sequence. Then, one decoder is applied to build the regression head for predicting future poses from the general representation. Finally, the other decoder is applied to build the classification head for predicting action semantics very early. Below, we first describe the detailed network structure of these encoders and decoders, and then we show how our model jointly predicts future human motion and action semantics.

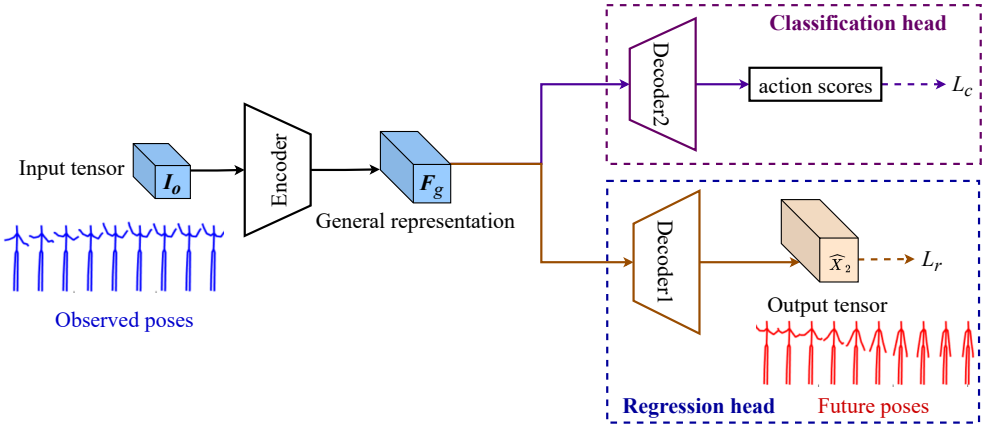



**Figure 3.** The architecture of the Multi-head TrajectoryCNN.

Encoders and decoders. As shown in Figure 4, (1) the encoder mainly consists of a convolutional layer and $n_1$ stacked RBs, and aims to extract a general spatiotemporal representation of the input sequence. (2) Decoder1 mainly consists of $n_2$ stacked RBs and two convolutional layers, and aims to reconstruct the position information of future poses from the general representation of the input sequence. (3) Decoder2 consists of $n_3$ stacked CBs, two fully connected (FC) layers, and a softmax layer, and aims to predict the semantic labels of current activities.

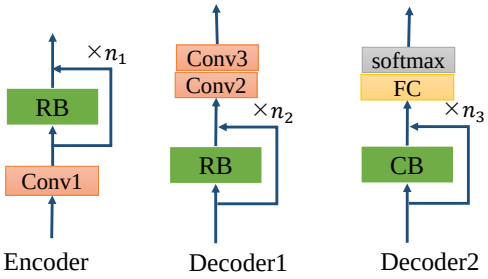



**Figure 4.** The structure of encoders and decoders. The orange boxes denote convolutional layers, where "Conv1" denotes a convolutional layer with a filter of $1 \times 1$ followed by a Leaky ReLU function, "Conv2" denotes a convolutional layer with a filter of $3 \times 3$ followed by a Leaky ReLU function, and "Conv3" denotes a convolutional layer with a filter of $1 \times 1$. The green boxes denote the regression block or the classification block in Section 3.1. The yellow box denotes the two fully connected layers. The gray box denotes the softmax layer. "$n$*" denotes the stacked numbers of RBs/CBs.

Human motion prediction. As shown in Figure 3, future human motion is predicted via the regression head, which is mainly adopted from [8]. Specifically, another decoder (i.e., Decoder1) is applied to predict future poses from the hidden representation of the inputs.

Action semantic prediction. As shown in Figure 3, the action semantics are predicted via the classification head. Specifically, we gather the general representation from the partial observations as the input, and the decoder built with CBs (i.e., Decoder2) is applied to obtain action scores of the future action. Instead of using RBs for the decoder, the proposed CBs are improved with MaxPooling layers to reach higher receptive fields for semantic predictions. From the theories of neural networks, the regression head and classification head are connected during backward propagation, which may be enhanced by each other for better performance.

### 3.3. Loss

As shown in Figure 3, our loss $L$ consists of two parts: pose regression loss $L_r$ and semantic prediction loss $L_c$, which can be formulated as Equations (2)–(4).

$$L = \alpha L_r + L_c \tag{2}$$

$$L_r = \left\| \mathbf{X}_2 - \widehat{\mathbf{X}}_2 \right\|_1 \tag{3}$$

$$L_c = CE(y, y') \tag{4}$$

where $\alpha$ is a scale factor, $\mathbf{X}_2$ denotes ground truth, $\widehat{\mathbf{X}}_2$ denotes the predicted future poses, $\|\cdot\|_1$ denotes $L_1$ norm loss, $y$ denotes ground truth, $y'$ denotes the predicted action semantic label, and $CE(\cdot)$ denotes the standard cross-entropy loss.

## 4. Experiments

### 4.1. Dataset and Implementation Settings

Dataset. Our method was evaluated on the NTU RGB+D dataset [42]. NTU RGB+D was collected by Microsoft Kinect v2 sensors, consisting of depth, 3D joints, RGB, and IR modalities. In this paper, we only use the 3D joint modalities. The dataset consists of 60 action classes and 56,880 samples collected by 40 subjects and cameras from different views, and each action class contains around 940 samples. The 40 subjects show varieties of age (i.e., 10∼35), gender, and height. The views contain three horizontal angles: $-45°$ (i.e., Camera 1), $+0°$ (i.e., Camera 2), and $+45°$ (i.e., Camera 3). To test the robustness of the algorithms to scale and perspective, two official protocols were adopted for evaluations, respectively, i.e., cross-subject evaluation and cross-view evaluation. (1) For the cross-subject evaluation, the training set comprised the samples of subject IDs 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38. The testing set comprised the samples of the remaining subjects. As a result, 40,320 and 16,560 samples were selected for training and testing sets, respectively. (2) For the cross-view evaluation, the samples of Camera 1 were used for testing, and the remaining samples were for training. As a result, 37,920 and 18,960 samples were selected for training and testing sets, respectively. For more details, refer to [42]. To efficiently evaluate the effectiveness of our method for both action semantic and human motion prediction, the lengths of all samples were processed to a fixed length of 64 frames by the uniform sampling strategies.

Implementation settings. All models were implemented with TensorFlow. Our learning rate was set to 0.0001 and the batch size was set to 256. All models were conducted on one GeForce GTX 1080 GPU. In experiments, the sequences on NTU RGB+D were unified sampled to a fixed length of 64. The lengths of partially observed motion sequence $\mathbf{X}_1$ (i.e., $N$) and future motion sequence $\mathbf{X}_2$ (i.e., $M$) were set to 32 (i.e., $N = M = 32$), and scale factor $\alpha$ was set to 0.01 empirically. $n$* was set to 2 in the encoders and decoders. The channels of Encoder1 were set to 256. The channels of RB in Decoder1 were set to 256, and the channels of the convolutional layers in Decoder2 were set to 32 (i.e., $M$). The channels of CBs in Decoder2 were set to 256 and 128 and the channels of FCs in Decoder2 were set to 256 and 60 (i.e., the action classes). Accuracy and mean per-joint position error (MPJPE) are introduced for our metrics. Accuracy is used for the evaluation of action semantic predictions, which indicates the percentage of correctly predicted categories. The MPJPE $E$ is used for the evaluation of future motion predictions, which can be formulated by Equation (5).

$$E = \frac{1}{J \times M} \sum_{n=1}^{M} \sum_{k=1}^{J} \left\| j_{nk} - \widehat{j}_{nk} \right\|_2 \tag{5}$$

where $M$ is the length of the future motion sequence, $J$ is the number of joints of the human body, $\widehat{j}_{nk}$ denotes the $k$-th predicted joint at the $n$-th pose, $j_{nk}$ denotes the ground truth joint, and $\|\cdot\|_2$ denotes the $L_2$ norm loss.

### *4.2. Comparison with State-of-the-Art*

Baselines. TrajectoryCNN and TrajectoryCNN-classification were selected as our baselines, among which "TrajectoryCNN" is for human motion prediction proposed in [8], and "TrajectoryCNN-classification" was modified based on TrajectoryCNN by replacing the convolutional layers in the decoder with two fully connected layers for predicting action semantics.

Results. Table 1 shows the results compared with the baselines. The "Accuracy" column denotes the performance of action semantic prediction, and the "Average error" column denotes the average error across all time steps for human motion prediction. Compared with "TrajectoryCNN", our proposed method obtained the lowest prediction errors of human motion by significant margins of +10.12 and +11.07 for cross-view (CV) and cross-subject (CS) evaluations, respectively. Compared with "TrajectoryCNN-classification", the accuracy of our method increased by +6.81% and +3.6% for CV and CS evaluations, respectively. Our method achieved the best performance of human motion prediction and action semantic prediction via a unified model, showing the effectiveness of our multi-task framework. During backward propagation, the regression head and classification head are connected, and their tasks can be enhanced by each other. Therefore, our method thus obtains an improved performance of human motion prediction and action semantic predictions. The results also reveal these facts: (1) action semantic information can significantly help human motion prediction. (2) Human motion prediction also helps to predict action semantics. (3) The action semantics and future human motion can be integrated to improve the performance of each other.

**Table 1.** Experimental results on NTU RGB+D, where "Average errors" denotes the average errors of MPJPE across all timestamps, "CV" denotes cross-view evaluation, and "CS" denotes cross-subject evaluation. The bold denotes the best results, and "+" denotes the improved performance of our method.

| Model | CV | | CS | |
|---|---|---|---|---|
| | **Accuracy** | **Average Error** | **Accuracy** | **Average Error** |
| TrajectoryCNN | - | 80.97 | - | 91.78 |
| TrajectoryCNN-classification | 67.74% | - | 60.22% | - |
| Multi-head TrajectoryCNN (Ours) | **74.55% (+6.81%)** | **70.85 (+10.12)** | **63.82% (+3.6%)** | **80.71 (+11.07)** |

Figures 5 and 6 show more detailed results quantitatively and qualitatively. As shown in Figure 5, our proposed method obtained the lowest errors at all timestamps, especially at the later timestamps, again showing the effectiveness of our method. Similarly, as denoted in Figure 6, our method still achieved more reasonable predictions, especially in the later predictions. Our superior results in Figures 5 and 6 largely benefit from our multi-task framework that jointly models human motion and action semantics, which encourages the network to predict the human motion of the activities.

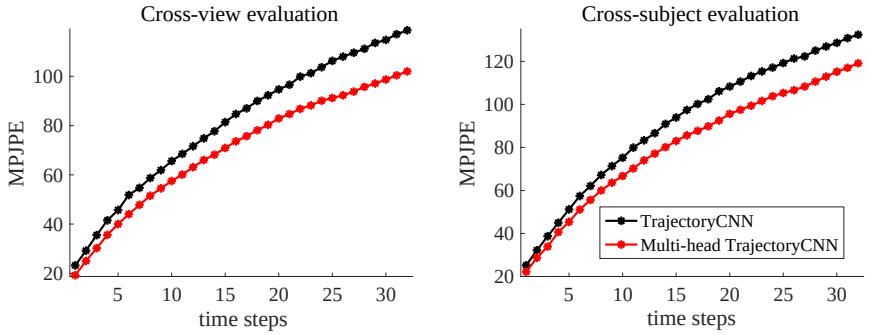



**Figure 5.** Frame-wise performance of human motion prediction. "MPJPE" denotes the mean per-joint position error of future poses by Equation (5), and "time steps" denotes the frames of future poses.

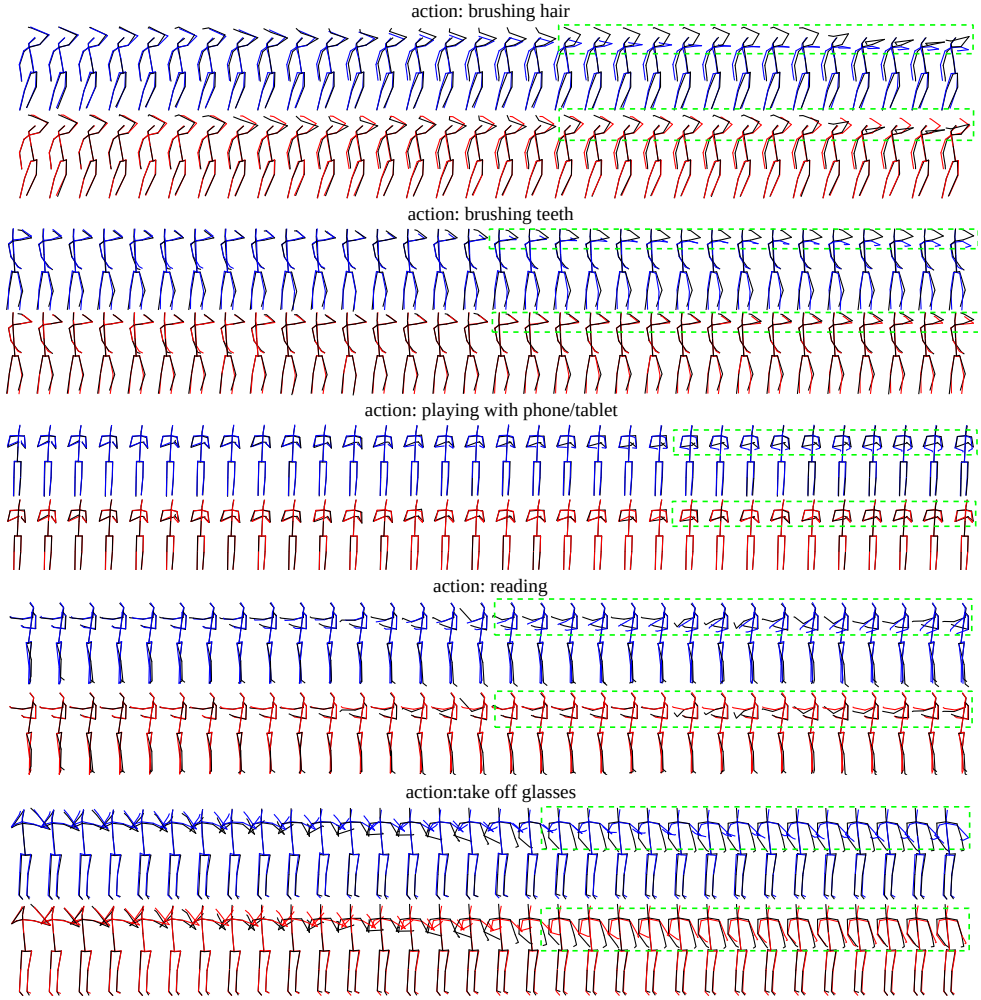



**Figure 6.** Visualization of future poses. The dark poses denote the ground truth, the blue poses denote the predictions of TrajectoryCNN [8], and the red poses denote the predictions of our method.

### *4.3. Ablative Study*

We will show the effectiveness of our proposed classification block (CB), the regression head (RH), and the classification head (CH). This section provides experimental results on the NTU RGB+D dataset using cross-view evaluation protocols. Detailed results are shown in Table 2. The "Accuracy" column denotes the performance of action semantic prediction, and the "Average error" column denotes the average error across all time steps for human motion prediction.

**Table 2.** Ablative results, where the bold denotes the best results.

| Model | Accuracy | Average Error |
|---|---|---|
| w/o CB | 69.44% | **70.64** |
| w/o RH | 66.11% | - |
| w/o CH | - | 71.15 |
| Multi-head TrajectoryCNN | **74.55%** | 70.85 |

Evaluation of CB. "w/o CB" shows the effectiveness of the classification block by replacing the CB with RB. Compared with "Multi-head TrajectoryCNN", the accuracy of action semantic predictions decreases by 5.11%, and the average errors are slightly improved by a margin of 0.21. Compared to the structures between "w/o CB" and "Multi-head TrajetcoryCNN", the main difference lies in the classification head. This may be the possible reason why the motion prediction is almost unchanged. The CB block is improved from RB by adding MaxPooling layers, which increases the receptive field to obtain higher semantic features for predictions. This may be the reason for the decreasing performance of action semantic predictions.

Evaluation of regression head. "w/o RH" shows the effectiveness of the regression head by removing the regression head in Figure 3. In this way, the network only focuses on predicting action semantics. Therefore, the results for motion prediction are not available. Table 2 only reports the accuracy of "w/o RH" for action semantic predictions. Compared with "Multi-head TrajectoryCNN", the accuracy of "w/o RH" decreases by 8.44%, showing that the regression head of future motion prediction can guide accurate predictions of action semantics.

Evaluation of classification head. "w/o CH" shows the effectiveness of the classification head by removing the classification head in Figure 3. In this way, the network only focuses on predicting future human motion while ignoring the prediction of action semantics. Therefore, the results for action semantics are not available. Table 2 only reports the average errors of "w/o CH" for human motion predictions. Compared with "Multi-head TrajectoryCNN", the errors of "w/o CH" increase by 0.3. The results show that the classification head's semantic predictions can also assist in predicting future human motion to some extent.

## 5. Conclusions

This paper has proposed a new multi-task framework, Multi-head TrajectoryCNN, for action prediction. In contrast to most prior works, it simultaneously predicts future human motion and action semantics via a unified framework. Specifically, we first proposed two building blocks for our network, including a regression block and a classification block. The regression block is used to extract general spatiotemporal features of the input sequence and build the regression head to predict future human motion. The classification block is improved from the regression head, enabling the model to capture higher semantic features for predictions of action semantics. Experiments on NTU RGB+D have shown the effectiveness of the proposed Multi-head TrajectoryCNN. The results also show that future human motion prediction can greatly help predict action semantics and semantic predictions also help predict future motion to some extent. Although we have achieved great performance, the proposed method suffers a limitation of missing some spatial characteristics of the human body. The natural connections of the human body can be represented by a graph, which can be processed by graph neural networks. We will explore this for efficiency and effectiveness in the future.

**Author Contributions:** Conceptualization, X.L. and J.Y.; methodology, X.L.; software, X.L.; validation, J.Y. and X.L.; formal analysis, X.L.; investigation, X.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L. and J.Y.; visualization, X.L.; supervision, J.Y.; funding acquisition, J.Y. and X.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported partly by the National Natural Science Foundation of China (Grant No. 62173045, 61673192), partly by the Fundamental Research Funds for the Central Universities (Grant No. 2020XD-A04-2), and partly by the BUPT Excellent Ph.D. Students Foundation (CX2021314).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The research in this paper used the NTU RGB+D action recognition dataset made available by the ROSE Lab at the Nanyang Technological University, Singapore.

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

## References

1. Zhao, H.; Wildes, R.P. Review of Video Predictive Understanding: Early Action Recognition and Future Action Prediction. *arXiv* **2021**, arXiv:2107.05140.
2. Barsoum, E.; Kender, J.; Liu, Z. Hp-gan: Probabilistic 3d human motion prediction via gan. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018.
3. Kundu, J.N.; Gor, M.; Babu, R.V. Bihmp-gan: Bidirectional 3d human motion prediction gan. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
4. Butepage, J.; Black, M.J.; Kragic, D.; Kjellstrom, H. Deep representation learning for human motion prediction and classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
5. Li, M.; Chen, S.; Zhao, Y.; Zhang, Y.; Wang, Y.; Tian, Q. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
6. Sun, C.; Shrivastava, A.; Vondrick, C.; Sukthankar, R.; Murphy, K.; Schmid, C. Relational action forecasting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
7. Wang, X.; Hu, J.F.; Lai, J.; Zhang, J.; Zheng, W.S. Progressive teacher-student learning for early action prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
8. Liu, X.; Yin, J.; Liu, J.; Ding, P.; Liu, J.; Liu, H. Trajectorycnn: A new spatio-temporal feature learning network for human motion prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2133–2146. [CrossRef]
9. Fragkiadaki, K.; Levine, S.; Felsen, P.; Malik, J. Recurrent network models for human dynamics. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
10. Martinez, J.; Black, M.J.; Romero, J. On human motion prediction using recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017.
11. Gui, L.Y.; Wang, Y.X.; Ramanan, D.; Moura, J.M.F. Few-shot human motion prediction via meta-learning. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
12. Chiu, H.; Adeli, E.; Wang, B.; Huang, D.A.; Niebles, J.C. Action-agnostic human pose forecasting. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Honolulu, HI, USA, 7–11 January 2019.
13. Jain, A.; Zamir, A.R.; Savarese, S.; Saxena, A. Structural-rnn: Deep learning on spatio-temporal graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
14. Wang, H.; Dong, J.; Cheng, B.; Feng, J. PVRED: A Position-Velocity Recurrent Encoder-Decoder for Human Motion Prediction. *IEEE Trans. Image Process.* **2021**, *30*, 6096–6106. [CrossRef] [PubMed]
15. Pavllo, D.; Feichtenhofer, C.; Auli, M.; Grangier, D. Modeling human motion with quaternion-based neural networks. *Int. J. Comput. Vis.* **2020**, *128*, 855–872. [CrossRef]
16. Li, Y.; Wang, Z.; Yang, X.; Wang, M.; Poiana, S.; Chaudhry, E.; Zhang, J. Efficient convolutional hierarchical autoencoder for human motion prediction. *Vis. Comput.* **2019**, *35*, 1143–1156. [CrossRef]
17. Li, C.; Zhang, Z.; Lee, W.S.; Lee, G.H. Convolutional sequence to sequence model for human dynamics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
18. Mao, W.; Liu, M.; Salzmann, M. History repeats itself: Human motion prediction via motion attention. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
19. Mao, W.; Liu, M.; Salzmann, M.; Li, H. Learning trajectory dependencies for human motion prediction. In Proceedings of the IEEE International Conference on Computer Vision 2019, Seoul, Korea, 27–28 October 2019.
20. Cui, Q.; Sun, H.; Yang, F. Learning dynamic relationships for 3d human motion prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
21. Kong, Y.; Gao, S.; Sun, B.; Fu, Y. Action prediction from videos via memorizing hard-to-predict samples. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

22. Hu, J.F.; Zheng, W.S.; Ma, L.; Wang, G.; Lai, J.; Zhang, J. Early action prediction by soft regression. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2568–2583. [CrossRef] [PubMed]
23. Sadegh Aliakbarian, M.; Sadat Saleh, F.; Salzmann, M.; Fernando, B. Encouraging lstms to anticipate actions very early. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
24. Chen, L.; Lu, J.; Song, Z.; Zhou, J. Part-activated deep reinforcement learning for action prediction. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
25. Wu, X.; Wang, R.; Hou, J.; Lin, H.; Luo, J. Spatial–temporal relation reasoning for action prediction in videos. *Int. J. Comput. Vis.* **2021**, *129*, 1484–1505. [CrossRef]
26. Chen, J.; Bao, W.; Kong, Y. Group activity prediction with sequential relational anticipation model. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
27. Wang, B.; Huang, L.; Hoai, M. Active vision for early recognition of human actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 16 June 2020.
28. Gammulle, H.; Denman, S.; Sridharan, S.; Fookes, C. Predicting the future: A jointly learnt model for action anticipation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
29. Pang, G.; Wang, X.; Hu, J.; Hu, J.F.; Zhang, Q.; Zheng, W.S. DBDNet: Learning Bi-directional Dynamics for Early Action Prediction. In Proceedings of the International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019.
30. Fernando, B.; Herath, S. Anticipating human actions by correlating past with the future with Jaccard similarity measures. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
31. Hu, J.F.; Zheng, W.S.; Ma, L.; Wang, G.; Lai, J. Real-time RGB-D activity prediction by soft regression. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
32. Kong, Y.; Tao, Z.; Fu, Y. Deep sequential context networks for action prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
33. Cai, Y.; Li, H.; Hu, J.F.; Zheng, W.S. Action knowledge transfer for action prediction with partial videos. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
34. Cai, H.; Bai, C.; Tai, Y.W.; Tang, C.K. Deep video generation, prediction and completion of human action sequences. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
35. Mahmud, T.; Hasan, M.; Roy-Chowdhury, A.K. Joint prediction of activity labels and starting times in untrimmed videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
36. Zhao, H.; Wildes, R.P. On diverse asynchronous activity anticipation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
37. Liao, D.; Liu, W.; Zhong, Y.; Li, J.; Wang, G. Predicting Activity and Location with Multi-task Context Aware Recurrent Neural Network. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018.
38. Guan, J.; Yuan, Y.; Kitani, K.M.; Rhinehart, N. Generative hybrid representations for activity forecasting with no-regret learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 16 June 2020.
39. Liu, M.; Tang, S.; Li, Y.; Rehg, J.M. Forecasting human-object interaction: Joint prediction of motor attention and actions in first person video. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
40. Li, B.; Tian, J.; Zhang, Z.; Feng, H.; Li, X. Multitask non-autoregressive model for human motion prediction. *IEEE Trans. Image Process.* **2020**, *30*, 2562–2574. [CrossRef] [PubMed]
41. Li, M.; Chen, S.; Chen, X.; Zhang, Y.; Wang, Y.; Tian, Q. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, *in press*. [CrossRef] [PubMed]
42. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.