# Multi-person Point Cloud Detection Transformer for Multi-Person Motion Classification

Yue Lin , Mingyang Liu , Jin Shi
Zhejiang University

## Abstract

*Point cloud videos provide a rich source for visual information, and also make it possible to recognize different actions more precisely in human agent interation. The ability to correctly recognize what people are doing is of great importance for robots to get involved in human life. A point cloud video is a sequence of frames each of which consists of a set of points with 3D coordinates. Because the points in frames are unordered and irregular, it is almost impossible to track each point from the start of the video to the end. Besides, different from single-person motion detection, multi-person(here we consider 2-person) motion detection requires splitting the points into two sets, one of a participant and another of the other participant. In this paper, in order to correctly classify the motions of two people, we propose a novel **Mu**lti-person **P**oint **C**loud **D**etection Trans**former** (MuPCDFormer). First, we develop a Bi-LSTM to split the point sets to several parts, each of which represents one of the two participants. Then, a convolution layer is developed to get spatial features from the point clouds. Finally, we develop a transformer to find out temporal features. We evaluate our framework on SBU Kinect Interaction Dataset (2 persons), and demonstrate that our method outperforms the state-of-the-art methods.*

## 1. Introduction

The field of robotics has seen significant advancements in recent years, enabling robots to play a more active and integral role in human life. A key aspect in achieving seamless human-robot interaction is the ability to accurately recognize and understand human actions. Point cloud videos have emerged as a valuable resource for extracting visual information, presenting a novel approach to enhance action recognition in human-agent interactions.

A point cloud video is a dynamic sequence of frames, where each frame comprises a collection of 3D points with their respective coordinates. This representation offers a unique challenge in tracking the movement of individual points throughout the video due to their unordered and irregular nature. Consequently, achieving precise and reliable action recognition in point cloud videos becomes a complex task, especially in multi-person scenarios where the points must be segregated into different sets, each corresponding to a specific participant.

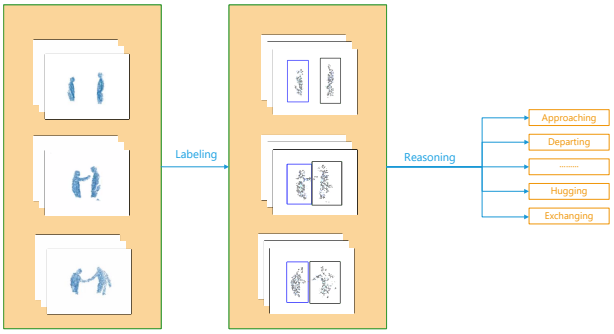


Figure 1. Overview of our task. First downsampling and labeling the point cloud videos into two parts, each of which represents one of the two participants. Then a series of networks are developed to reason the type of the multi-person action. See introduction (§1) for details.

In order to understand dynamics in point cloud videos, one solution is to first convert point cloud videos into a sequence of regular, ordered voxels and then apply grid based convolution to these voxels. However, as points are usually sparse, special engineering efforts, *e.g.*, sparse convolution [3] are usually needed. Besides, the voxelization process may lose some information and require additional computation [24].

In this paper, we developed a person-aware transformer-based framework for multi-person action classification in point cloud videos free of voxelization. The overview task of our model is explained in 1. Our framework is composed of two procedures: First downsample and label each point in each frame to different persons, and then go through a series of neural network to reason the type of the motion input. Note that labeling points into separate persons can help the model better understand person-specific features and movements. We evaluate our framework on SBU Kinect Interaction Dataset [28], and demonstrate that our method outperforms the state-of-the-art methods.

In general, our contributions are as follows: 1) We propose a novel framework for multi-person action classification in point cloud videos. To our knowledge, this is the first work to consider the multi-person action classification in point cloud videos. 2) We adopt a sublayer of our network to split the point cloud videos into two parts, each of which represents one of the two participants. This is important for multi-person action classification. 3) Our proposed model outperforms the state-of-the-art methods on SBU Kinect Interaction Dataset.

## 2. Related Work

### 2.1. Point Cloud Detection.

**Spatial-Temporal Modeling in Grid based Videos.** Deep neural network has achieved great performance on spatio-temporal modeling in RGB/RGBD videos. Since video is a kind of sequence, recurrent newral networks [4, 20, 29] are used to capture the temporal dependencies [7, 27]. Meanwhile, 3D convolution neural networks can learn spatio-temporal representations from videos.

**Deep Learning on Static Point Clouds.** Deep learning has been widely used in a lot of point cloud problems, for instance, classification, scene semantic segmentation , reconstruction [6, 15, 26] and object detection. These methods mainly focus on static point clouds rather than consider the temporal dynamics of point clouds.

**Point Cloud Video Processing.** Point cloud video modeling is a fairly new task but is of great importance for intelligent agents to understand the dynamic 3D world we live in. MinkowskiNet [3] uses 4D Spatio-Temporal ConvNets to extract appearance and motion from 4D occupancy voxel grids. PSTNet [9] constucts the spatio-temporal hierarchy to alleviate the requirement of point tracking. A method by 3DV [24], is developed which first employs a temporal rank pooling to merge point motion into a voxel set and then applies PointNet++ [19] to extract the spatio-temporal representation from the set. P4Transformer [8] embeds the spatio-temporal local structures presented in a point cloud video and capture the appearance and motion information accoress the entire video by performing self-attention on the embedded local features.

### 2.2. Human-Human Interaction Understanding.

In order to understand human-human interactions, several approaches, mainly about pose and trajectory forecasting, are proposed. For example, Wang et al. [23] present a Transformer-based framework to forecast multi-person motion. Furthermore, Peng et al. [18] learns skeletal body part dynamics between intra- and inter-individuals to better predict multi-person's motions. Besides, Chopin et al. [2] propose a Transformer based architecture for the human reaction generation task. Pedestrian trajectory prediction is also a representative issue for multi-person social interaction, where RNNs [20], Transformer [21] and Graph neural networks(GNNs) [13] are often adopted as social models [1, 11, 12, 14, 25] for interaction modeling. While performing well, these studies only focus on skeletal body. In this work, we investigate our MuPCDFormer to consider interactions in the scale of point cloud videos, and reason their categories of each interaction.

## 3. Method

In this section, we introduce our Multi-Person Point Cloud Detection Transformer (MuPCDFormer) for multi-person action classification, which contains a Bi-LSTM module, a convolution module and a Transformer module followed by fully connected layers, as shown in Figure 2. In the following, the problem definition and the details of each module are described in detail.

### 3.1. Problem Definition

Supposing the observed point clouds from the people are $X_{1:T} = \{X_1, X_2, \cdots, X_T\} \in \mathbb{R}^{n\times 3}$ with T frames, where $X_i$ contains $n = 4096$ points with xyz coordinates. Given the motions of the people, our goal is to get the type of multi-person actions.

### 3.2. Bi-LSTM

After downsampling the point by farthest point sampling [19], we use Bi-LSTM to label each point in each frame to one of the participants. Bi-LSTM can gather information between frames, especially, it can obtain the information of both the start and end frame. This is quite important because actions like approaching, points can be easily distinguished at the start frame, while actions like departing, points can be easily distinguished at the end frame. Although points are orderless between frames, we believe that our model can figure out the underlying relation between frames, and we prove that this procedure helps the model perform better in ablation study (§4.5). The Bi-LSTM can be represented as

$$
\begin{aligned}
i_t &= \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \\
f_t &= \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \\
o_t &= \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \\
\tilde{c}_t &= \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh\left(c_t\right)
\end{aligned}
$$

where $x_t$ is the input at time $t$, $h_t$ is the hidden state at time $t$, $c_t$ is the cell state at time $t$, $i_t$, $f_t$, $o_t$ are the input gate, forget gate and output gate at time $t$, respectively. $W_i, W_f, W_o, W_c$ are the weight matrices for input, forget, output and cell, respectively. $U_i, U_f, U_o, U_c$ are the weight matrices for input, forget, output and cell, respectively. $b_i, b_f, b_o, b_c$ are the bias vectors for input, forget,

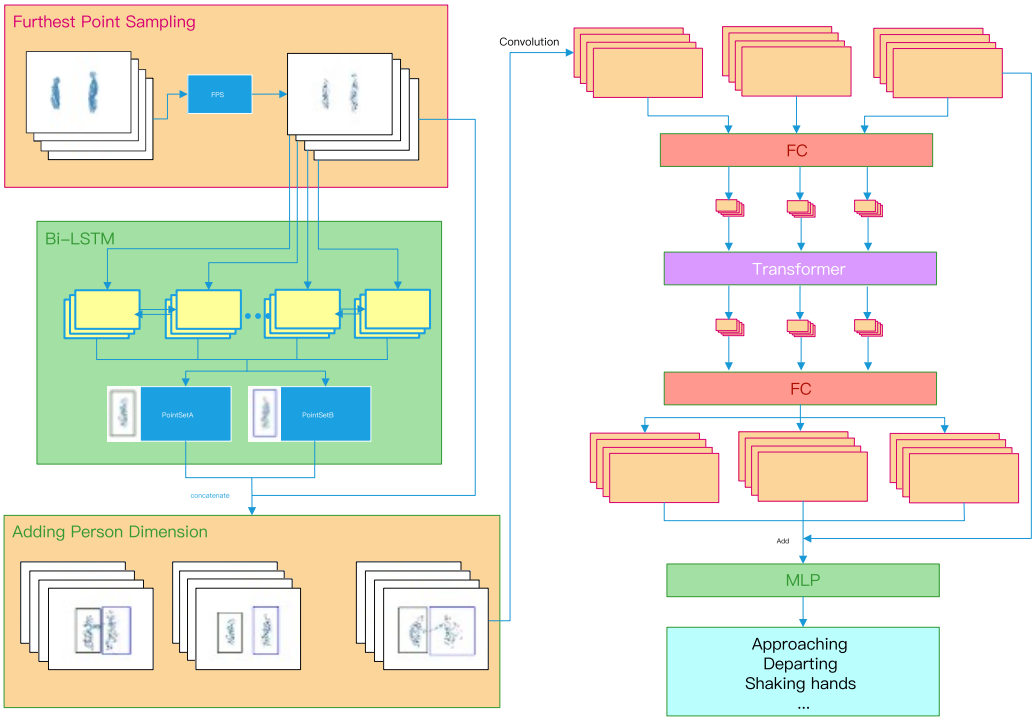


Figure 2. **Framework** of our proposed MuPCDFormer. Given the input point cloud videos, MuPCDFormer downsample them by Farthest Point Sampling [19], and then feed them into Bi-LSTM module to split the point cloud into two parts, each of which represents one of the two participants. Then, a convolution layer is developed to get spatial features from the point clouds. Finally, we develop a transformer to find out temporal features. See method (§3) for details.

output and cell, respectively. $\sigma$ is the sigmoid function. $\odot$ is the element-wise multiplication.

### 3.3. Convolution

After Bi-LSTM which gains the person information, we concatenate it with $X_i$ to obtain a new point cloud $X_i' \in \mathbb{R}^{n\times 4}$ which contains the person information. Then we use a series of convolution layer to get spatial features from the point clouds. The convolution layers extract features from two dimensional $X_i'$ to one dimensional $y_i \in \mathbb{R}^{1024}$, which can be represented as

$$y_i = \sigma\left(W_i X_i' + b_i\right)$$

where $x_i$ is the input at time $i$, $y_i$ is the output at time $i$, $W_i$ is the weight matrix, $b_i$ is the bias vector, $\sigma$ is the relu function.

### 3.4. Transformer

To better retain the temporal information of the point cloud videos, we use Transformer to capture the temporal information of the point cloud videos. In order to let the input dimensional not so large, we deploy two fully connected layers between and after Transformer. The transformer takes as input queries $Q$, keys $K$, and values $V$, each of which is projected by the corresponding parameter matrix $W^Q$, $W^K$, $W^V$. The output of Transformer is computed as

$$Q = y_i W^Q, K = y_i W^K, V = y_i W^V$$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Then we use a feed forward network to get the output of the transformer, and add it with $y_i$ like residual connection. Finally we use a fully connected layer and softmax function to generate the predicted distribution of the type of the multi-person action.

### 3.5. Loss Function

Here we use Cross Entropy Loss, which is common for multi-classification problems. In particular, the loss is represented as

$$H(p, q) = -\sum_x p(x) \log q(x)$$

where $p$ is the ground truth distribution and $q$ is the predicted distribution. In our case, $p$ is the one-hot vector of the ground truth label and $q$ is the softmax output of the network.

| Method | Accuracy |
|---|---|
| P4Transformer [8] | 0.754 |
| PointNet++ [19] | 0.661 |
| Ours | **0.799** |

Table 1. **Results** (§4.4) of accuracy on SBU dataset. We compare our method with previous SOTA methods on classification of multi-person action. Best results are in boldface.

# 4. Experiments

## 4.1. Implementation Details

We implement our framework in PyTorch, and the experiments are performed on Nvidia GeForce RTX 2080Ti GPU. We train our model for 3000 epochs using the ADAM optimizer with a batch size of 16, a learning rate of 0.0001, which decay 50% every 20 epoch but jump to original learning rate (0.0001) every 50 epoch. This may help jump out of local optima and head for final optima. For regularization, a weight decay rate of 0.0001 is deployed. When downsampling at first, we choose 512 anchor point. For Transformer, we choose a dimensionality of 32 with 2 stacked encoder/decoder and 8 head attention.

## 4.2. Dataset

**SBU Dataset [28]** contains 8 classes of simple interaction motions: walking toward, walking away, kicking, pushing, shaking hands, hugging, exchanging, and punching. To unify the frames between different sequences and point numbers in each frames, we choose 4096 points each frame, which is double of [8] for single person action classification, and 21 frames per video which is the minimum frames among sequences. We follow the splits in [2], which leaves us 1188 seqs for training and 224 seqs for testing.

## 4.3. Baselines

We choose two code-released approaches as the baselines, both are single-person based methods. The first is P4Transformer [8], we trained the model with 21 frames of the input and let the model predict the type of the two-person action. Another baseline is PointNet++ [19] which is quite similar to ablation on LSTM&Transformer, so we reproduce their code and train it with SBU dataset.

## 4.4. Results

In order to validate the effectiveness of our method, we compare our method with the baselines. The results are shown in Table 1. We can see that our method outperforms the baselines discussed before. The reason is that our method can capture the spatio-temporal information of the point cloud videos as well as person information, which is important for multi-person action classification.

| Method | Accuracy |
|---|---|
| w/o Bi-LSTM | 0.759 |
| w/o Transformer | 0.719 |
| w/o Bi-LSTM & Transformer | 0.661 |
| Full | **0.799** |

Table 2. **Ablation Studies** (§4.5) on different components of MuPCDFormer. Our full method and its variants are evalutaed on SBU dataset.

## 4.5. Ablation Study

We conduct ablation study on our model to verify the effectiveness of each component. The results are shown in Table 2.

**Effectiveness of Bi-LSTM.** We first remove the Bi-LSTM module from our model, and the accuracy drops from 0.799 to 0.759. This is because the Bi-LSTM module can help split the point cloud into two parts, each of which represents one of the two participants. Assigning points to individuals helps the model to capture person-specific features and movements, hence improving the performance of the model.

**Effectiveness of Transformer.** We then remove the Transformer module from our model, and the accuracy drops from 0.799 to 0.719. This is because the Transformer module can help capture the temporal information of the point cloud videos, which is important for multi-person action classification.

**Effectiveness of Bi-LSTM & Transformer.** We finally remove both the Bi-LSTM and Transformer modules from our model, and the accuracy drops from 0.799 to 0.661. This is because both the Bi-LSTM and Transformer modules are important for multi-person action classification, without learning the temporal and person features of the point cloud videos, the model cannot perform well.

# 5. Conclusion

We introduce a novel framework for multi-person action classification in point cloud videos. Our framework can capture the spatio-temporal information of the point cloud videos as well as person information, which is important for multi-person action classification. We evaluate our framework on SBU Kinect Interaction Dataset, and demonstrate that our method outperforms the state-of-the-art methods.

# Appendix

## Problems Found and How to Solve

**Problem 1.** The first problem is that the point cloud videos are hard to obtain, especially for the videos with at least two people interacting with each other. We surveyed a range of multi-person interaction datasets 3, and found that most of them have only skeletal data and RGB data, but no point cloud data. We finally found the SBU Kinect Interaction Dataset [28] which contains depth map and can be converted into point cloud videos.

**Problem 2.** The second problem is how can we obtain point cloud videos from depth map. It seems that this process is an optical problem and can be solved by some optical methods, but we are not familiar with this field. We finally found a method from [16] which can convert depth map into point cloud videos.

**Problem 3.** The third problem is we encounter some training problem like loss not decreasing and CUDA out of memory. We finally found that the reason is that the point numbers of the point cloud videos are too large, and we need to downsample the point cloud videos to a reasonable size. It seems that the LSTM will take a large amount of memory that makes the CUDA out of memory and the loss not decreasing. The dimensionality of the transformer is also too large so the attention score is too high, after softmax, the gradient is almost zero, making the loss not decreasing.

**Problem 4.** The fourth problem is that we find it kind of hard to find the real optima. The learning rate should be set to around 0.0001 after several experiments, otherwise the loss will be difficult to decrease. Also, in order to find the real optima, we need to decay the learning rate to 50% every 20 epoch but jump to original learning rate (0.0001) every 50 epoch. This may help jump out of local optima and head for final optima.

| Name | Sequence | Types |
|---|---|---|
| SBU Kinect [28] | 225 | approaching, departing, kicking, pushing, shaking hands, hugging, exchanging, punching |
| K3HI [10] | 264 | approaching, departing, kicking, pushing, shaking hands, pointing, exchanging, punching |
| 3DPW [22] | 204 | walk/run, sitting, talking, shaking hands, dancing, hugging, cheering, going into cars, $\cdots$ |
| Mupots3d [17] | 99 | punching, sitting, walking, kicking, push-up, press-up, shaking hands, playing balls, jumping |
| CMU Mocap [5] | 58 | walking, shaking, dancing, comforting, sheltering, stumble into, throw & catch, sits and pull |

Table 3. **DataLists** (§5) on different components of MuPCD-Former. Our full method and its variants are evalutaed on SBU dataset.

## References


[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016. 2

[2] Baptiste Chopin, Hao Tang, Naima Otberdout, Mohamed Daoudi, and Nicu Sebe. Interaction transformer for human reaction generation. *IEEE Transactions on Multimedia*, pages 1–13, 2023. 2, 4

[3] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1, 2

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 2

[5] CMU-Graphics-Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003. 5

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[7] Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018. 2

[8] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021. 2, 4

[9] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. *arXiv preprint arXiv:2205.13713*, 2022. 2

[10] Tao Hu, Xinyan Zhu, Wei Guo, Kehua Su, et al. Efficient interaction recognition through positive action representation. *Mathematical Problems in Engineering*, 2013, 2013. 5

[11] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019. 2

[12] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. PMLR, 2018. 2

[13] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[14] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020. 2

[15] Ruihui Li, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7203–7212, 2019. 2

[16] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9246–9255, 2019. 5

[17] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 5

[18] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17121–17130, June 2023. 2

[19] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[20] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997. 2

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[22] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 601–617, 2018. 5

[23] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6036–6049. Curran Associates, Inc., 2021. 2

[24] Yancheng Wang, Yang Xiao, Fu Xiong, Wenxiang Jiang, Zhiguo Cao, Joey Tianyi Zhou, and Junsong Yuan. 3dv: 3d dynamic voxel for action recognition in depth video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 511–520, 2020. 1, 2

[25] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 507–523. Springer, 2020. 2

[26] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 2

[27] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[28] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 28–35. IEEE, 2012. 1, 4, 5

[29] Xiaohan Zhang, Lu Liu, Guodong Long, Jing Jiang, and Shenquan Liu. Episodic memory governs choices: An rnn-based reinforcement learning model for decision-making task. *Neural Networks*, 134:1–10, 2021. 2