WHEN SKELETON MEETS APPEARANCE: ADAPTIVE APPEARANCE INFORMATION ENHANCEMENT FOR SKELETON BASED ACTION RECOGNITION

Suqin Wang[1,2], *Lu Zhou*[1,2], *Yingying Chen*[1,2,3], *Jiangtao Huo*[4], *Jinqiao Wang*[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences,Beijing,China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences,Beijing,China
[3] Development Research Institute of Guangzhou Smart City
[4] Army Medical University, NCO School of PLA
wangsuqin2019@ia.ac.cn; {lu.zhou, yingying.chen, jqwang}@nlpr.ia.ac.cn; Jiangtaohuo@126.com.

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9859589

ABSTRACT

Skeleton-based action recognition methods which utilize graph convolution networks (GCNs) have achieved remarkable success in recent years. However, action recognizer can be easily confused by the ambiguity caused by different actions with similar skeleton sequences when only skeleton data is trained. Introducing appearance information can effectively eliminate the ambiguity. Based on this, we introduce a two-stream network for action recognition. One trained on RGB images extracts appearance information. The other trained on skeleton data models motion information and adaptively captures appearance information of action areas at action-related intervals via a specially tailored attention mechanism. Our architecture is trained and evaluated on two large-scale datasets: NTU RGB+D and NTU RGB+D 120, and a small scale human-object interaction dataset Northwestern-UCLA. Experiment results verify the effectiveness of our method and the performance of our method exceeds the state-of-the-art with a significant margin.

Index Terms— Action recognition, skeleton data, RGB images, attention.

1. INTRODUCTION

Action recognition has received a significant amount of attention in recent years, as it plays a significant role in a number of real-world applications. It is can be used in human-computer interaction, intelligent video surveillance, robot vision, etc.

With the development of different kinds of accurate and affordable sensors, multiple modalities are used for action recognition. Recent years have witnessed an emergence of works [1–3], using various data modalities for action recognition, such as RGB, skeleton, and multi-modality fusion. Among these modalities, skeleton data which encodes the trajectories of human body joints is succinct and efficient for action recognition and is robust to variations of clothing textures and backgrounds. Due to these advantages, skeleton-based action recognition methods have attracted much attention in the research community. More recently, many GCN-based action recognition methods [4, 5], which explore how to build a better topology and how to update the features of node have been proposed and achieved better performance. Nevertheless, according to our analysis of action recognition results based on Shift-GCN [5], we conclude that there is ambiguity in some actions with similar poses, i.e., reading, typing on a keyboard as shown in Fig. 1. To better recognize this kind of action, we need to add appearance information of the action-related areas for action recognition.

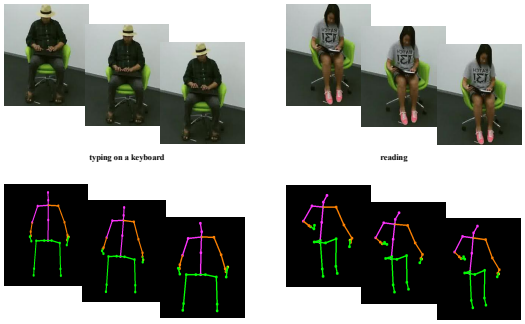

Fig. 1. Difficult action pairs with similar poses like 11 and 30 in NTU RGB+D dataset. The first line shows RGB images of typing on a keyboard and reading, and the second line shows the skeleton visual images of typing on a keyboard and reading.

In general, RGB data is easy to collect and contains abundant appearance information of the captured scene context. RGB-based action recognition methods [6, 7], have mostly learnt to exploit contextual information (e.g., scene class, dominant objects, and background motion). Whereas they rarely pay attention to understanding the human action itself. As a consequence, we need to add appearance information of

the areas associated with the action for action recognition.

Based on this, we propose a two-stream network to learn action features which include not only the specific features of dynamic skeleton modality but also the detailed appearance features of action-related areas at key intervals when the action occurs. Firstly, in order to obtain the appearance information of the action-related areas, we divide the human body into three partitions according to where the action takes place. Secondly, AFP (Appearance Features Process) module is used to process appearance features, and a tailored attention mechanism is used to select the features of the action-related areas in key intervals. Specifically, skeleton features are used as global information to generate corresponding weight values for different partitions in different times. At last, the dynamic skeleton information and the appearance information of the action-related areas in key intervals are used to determine the action category. The major contributions of this method lie in three aspects:

- According to the main areas where the action take place, we divide the human body into three partitions.
- We propose a customized attention mechanism which uses skeleton features as global information to calculate the response values of different body partitions at different times.
- On two human activity datasets and a human-object interaction dataset, the proposed model achieves superior performances compared to the earlier methods.

2. RELATED WORK

RGB. 3D CNN-Based methods simultaneously model the spatial and temporal context information in videos. Tran et al. [8] extended 2D convolution kernels to 3D convolution kernels to learn spatio-temporal features for action recognition. However, it brought a lot of parameters to train. To decrease the computational cost, [7,9] factorized 3D convolution to 2D spatial convolution layer followed by 1D temporal convolution layer. Similarly, Lin et al. [10] proposed a Temporal Shift Module (TSM), which shifts the channels along the temporal dimension both forward and backward, thus the information is exchanged between adjacent frames, and the complexity is maintained to the level of 2D CNNs.

Skeleton.With the development of deep learning, data-driven methods have become the mainstream methods. Method [11] based on CNN modelled the skeleton data as a pseudo-image, which manually design transformation rules. Method [12] based on RNN modelled the skeleton data as a sequence of coordinate vectors which represent human body joints. These methods failed to fully represent the structure of the skeleton data as the skeleton data are naturally embedded in the form of graphs. So there been have many works using graph convolution to process skeleton data. [4, 13] directly preformed the convolution filters on the graph vertexes and their neighbors.

RGB+Skeleton. Different modalities usually have distinct strengths and limitations for action recognition. It is an inevitable choice to fuse data of multiple modalities and take advantage of these advantages in action recognition. These modalities must be processed by different kinds of network to show their effectiveness, owing to they are heterogeneous. [14] proposed a new hierarchical bag-of-words feature fusion technique based on multi-view structured sparsity learning to fuse atomic features of two disparate modalities. [15] extracted spatial features from a middle frame using two attention modules, a self-attention and a skeleton-attention module. Temporal features are extracted from skeleton sequence by a BI-LSTM sub-network. The spatial features and the temporal features are combined for action recognition. These simple multi-modal fusion strategy limits their performance. As a result, many methods driven by skeleton data used attention mechanism to make RGB modality focus on the features of action. [16] used LSTM which is used to extract features from skeleton data to learn spatial and temporal attention weights, then the weights were multiplied with the feature map extracted from RGB data. VPN [17] had 2 key components which were an attention network and a spatial embedding.

But these methods have some flaws: 1) The attention module neglect the action is related to a small part of the human body. 2) These methods only end up relying on features from RGB videos for action recognition. Thus, they pay more attention to appearance information than dynamic of skeleton data.

3. METHODOLOGY

In this section, we explain the proposed two-stream network in detail. Firstly, we briefly present the overall pipeline of our proposed framework. Secondly, we describe the strategy of dividing the human body into three partitions based on where the action takes place. Finally, we introduce AFP module and tailored attention mechanism.

3.1. Pipeline Overview

In order to obtain skeleton features and appearance features, we devise a two-stream network architecture for action recognition, as shown in Fig.2. Firstly, we introduce the input and feature extraction backbone of the network. The inputs of our proposed model consist of the RGB images (randomly sampling 8 human body frames) and the joint coordinates. The skeleton data is extracted from motion-capture device. For RGB data, we use TSM [10] to extract the appearance representation f. f is a feature map of dimension $T \times W \times H \times C$, where T denotes the temporal dimension, $W \times H$ the spatial scale and C the channels. The skeleton data is processed

by Shift-GCN [5] to extract skeleton features g. The dimension of g is $1 \times D_1$. Secondly, we use the AFP module to process features f to generate partition features f'' and f'''. Features f'' and f''' serve as Q and V of the tailored attention mechanism. The dot product of skeleton features g' and features f'' is input into softmax function to learn the spatial and temporal attention weights of partitions. Finally, final appearance features are obtained by using average pooling to process weighted appearance features along the temporal dimension. We concatenate final appearance features and skeleton features g for action recognition. The final features include not only the pose information, but also the appearance information of the action-related areas in key intervals.

3.2. Human Body Partition

For confusing actions which have similar posture, we need to add appearance information of the areas associated with the action to assist skeleton data for action recognition. According to the areas where the action occurs, we can simply divide the human body into three partitions as shown in Fig.3. The exact location of each partition is determined by the coordinates of the upper left corner and the size of the partition. Firstly, according to the skeleton data, the human body is cut on the corresponding image. The upper left corner of the human body is the minimum value of the abscissa and the minimum value of the ordinate in all joints of the human body, and the lower right corner is the maximum value of the abscissa and the maximum value of the ordinate among all joints of the human body. Secondly, we locate the position of each partition. The three partitions are equal in length to the human body. For the first partition, the width is the ordinate of the neck minus the ordinate of the upper left corner of the human body and the upper-left coordinate is the upper-left coordinate of the human body. For the second partition, the width is the difference between the ordinate of the middle torso and the ordinate of the neck, and the upper-left coordinate is the same as the lower-left coordinate of the first partition. For the third partition, the width is the difference between the ordinate of the lower right corner of the human body and the ordinate of the middle torso, and the upper left coordinate is the same as the lower lower-left coordinate of the second partition.

3.3. AFP Module and Attention Mechanism

How should we deal with the appearance features? We use the AFP module to process f shown in Fig.3 to generate f'' and f'''. Firstly, the upper left corner coordinate, length and width of each partition obtained in Section 3.2 are used as the input of ROIPooling [18] to obtain the features f_i ($P \in R^{T \times W_1 \times H_1 \times C}$), where i $\in (1, 2, 3)$, of the corresponding ith partition at different times from the features f. Secondly, average pooling is used to process the features f_i to get features f_i'. Finally, we use two branches to process the features f_i' to obtain f_i'' and f_i'''. Features f_i' is followed by Fc Layer, a non-linear mapping function and Fc Layer to obtain f_i''. Here we choose tanh as the non-linear mapping function. These features are defined as:

$$f_i'' = Fc\left(\tanh\left(Fc\left(f_i'\right)\right)\right), \tag{1}$$

where Fc is the full connection layer. Features f_i' are followed by two Fc Layers to obtain f_i''', which can be defined as:

$$f_i''' = Fc\left(Fc\left(f_i'\right)\right), \tag{2}$$

Features f'' and f''' are obtained by the concatenation of f_i'' and f_i'''. For simplicity, the processing of the skeleton features g is quite simple and g is followed by only one Fc Layer and a non-linear mapping function:

$$\mathrm{g}' = \tanh(Fc(g)), \tag{3}$$

Next, we use the attention mechanism to adaptively select the appearance features of the areas associated with the action in key intervals, where features g' are used as K, features f'' are used as Q and f''' is used as V. We use the following formula to calculate the weights of the ith partitions at different times:

$$w_i = \frac{f_i'' \cdot g'}{\sum_{i=1}^{3} f_i'' \cdot g'}, \tag{4}$$

where w_i is the temporal weights of the ith partition. Afterwards, the features f_i''' are multiplied by its corresponding weights w_i to generate the weighted appearance features f_{wi} of the ith partition.

$$f_{wi} = w_i f_i'''. \tag{5}$$

So far, the weighted appearance features F' concatenating the weighted appearance features of the three partitions pay more attention on the relevant partitions where the action occurs in key intervals. At last, we concatenate the final appearance features F that average pooling is used to process to process F' on the temporal dimension and the skeleton features g to form the final features. It not only contain pose information but also detailed information about the areas where the action occurs in key intervals. Through the above processing, we take the final features as classifier input to recognize action.

4. EXPERIMENTS

4.1. Setup and Dataset

NTU RGB+D. NTU RGB+D [19] is a large-scale action recognition dataset containing 56,880 skeleton sequences and video samples, which are performed by 40 distinct subjects, captured from 3 different camera view angles and categorized into 60 classes. Each skeleton sequence contains 25 body joints. Each sample contains an action and is guaranteed

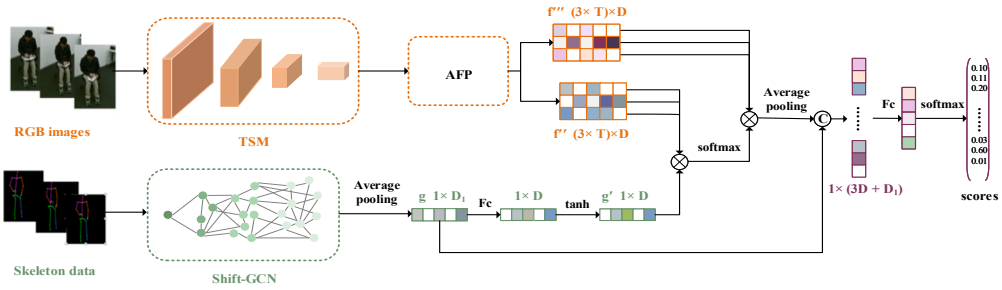

Fig. 2. The architecture of our proposed two-stream network. It consists of the RGB stream and the skeleton stream. The skeleton stream models dynamics of posture from skeleton sequences and the RGB stream extracts spatio-temporal appearance features from images. Posture motion features are used as global information to guide appearance features to select features of the action-related partition in key intervals.

Fig. 3. There are some examples of human-object interaction. The solid red boxes show the area where the action takes place, and the green dashed boxes represent three partitions.

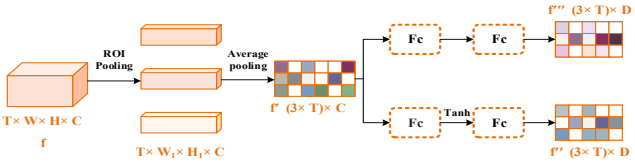

Fig. 4. The process method for appearance features extracted from RGB backbone. The components in process features: extraction of partition feature and partition feature processing.

to have at most 2 subjects, which is captured by three Microsoft Kinect v2 cameras from different views concurrently. The author of this dataset recommends two benchmarks: 1) cross-subject (X-sub) benchmark: the 40 subjects are split into training and testing data, training data is obtained from 20 subjects and testing data is obtained from the other 20 subjects. 2) cross-view (X-view) benchmark: the 3 views are split into training and testing data, training data is obtained from the camera views 2 and 3, and the testing data is obtained from the camera view 1.

NTU RGB+D 120. NTU RGB+D 120 [20] extends NTU RGB+D by adding another 57600 video samples and skeleton sequences, which are performed by 106 distinct subjects, captured from 32 setups that each setup denotes a specific location and background and categorized into 120 classes. This dataset is the largest dataset with 3D joints annotations for action recognition. The author of this dataset recommends two benchmarks: (1) cross-subject (X-sub) benchmark: the 106 subjects are split into training data and testing data containing 53 subjects respectively. (2) cross-setup (X-setup) benchmark: the 32 setups are split into training data with even setup IDs and testing data with odd setup IDs.

Northwestern-UCLA. Northwestern-UCLA [21] is captured by three Kinect cameras. It contains 1494 video clips covering 10 categories. Each action is performed by 10 actors. We adopt the same evaluation protocol in [18] : we use the samples from the first two cameras as training data and the samples from the other camera as testing data.

Experiment Settings. All models use SGD with momentum 0.9 when training the model. The total epochs is 60. Initial learning rate is set to 0.01 and dropped by 0.1 at epoch 20, 40. For NTU-RGB+D [19] and NTU-RGB+D 120 [20], the batch size is set to 8. We employ ImageNet pre-training when training TSM [10].

4.2. Ablation Study

In this subsection, we first show that appearance information can significantly improve the performance of model only based on skeleton data. Then we demonstrate the effectiveness of our method. In order to verify appearance information can assist skeleton data for action recognition, we concatenate features g and features f' named L1, where are concatenated by f'_1, f'_2 and f'_3. As shown in Table 1, concatenating multi-modality features can improve the classification of actions (upto 2.2% than TSM and 4.6% than Shift-GCN) on NTU RGB+D X-sub (CS) and (upto 2% than TSM and 0.7% than Shift-GCN) NTU RGB+D X-view (CV). As shown in Table 1, concatenating multi-modality features with ours is more effective (up to 1.7% on NTU RGB+D X-sub and 1.1% on NTU RGB+D X-view). This phenomenon indicates that concatenating all appearance features bring some misleading information. Our method pays more attention to the parti-

Fig. 5. The figure shows the weights of the different partitions obtained using our method in different times.

tions that can distinguish actions to reduce the interference of misleading information.

Table 1. Performance comparison of different fusion strategies on X-sub benchmark and X-view benchmark of NTU RGB+D.

Method	Att	CS	CV
TSM	×	90.12	93.55
Shift-GCN	×	87.71	94.82
L1	×	92.30	95.59
ours	√	94.02	96.68

4.3. Comparison with The State-of-the-art

We compare our method with the state-of-the-art on NTU RGB+D, NTU RGB+D 120, and N-UCLA in Table 2, 3, and 4. The performance of our method is superior to other methods in three datasets. In Table 2, for input modality RGB+skeleton, our method improves the state-of-the-art by up to 0.5% on NTU RGB+D CS and CV. To make a fair comparison, VPN [10] uses I3D as the backbone to process RGB data, so we replace TSM with I3D. The performance is 96.50% on CV. Compared with single mode, the performance is also improved, which proves the effectiveness of our method.

Table 2. Comparisons of the validation accuracy with state-of-the-art methods on the NTU RGB+D dataset.

Method	Ske.	RGB	Att	CS	CV
STA-Hands	√	√	√	82.5	88.6
altered STA-Hands	√	√	√	84.8	90.6
PEM	√	√	×	91.7	95.2
Separable STA	√	√	√	92.2	94.6
P-I3D	√	√	√	93.0	95.4
VPN	√	√	√	93.5	96.2
ours	√	√	√	94.0	96.7

Compared to the state-of-the-art results, the improvements of 1.5% and 0.5% on NTU RGB+D 120 X-sub (CS_1) and X-setsub (CS_2) respectively are significant as shown in Table 3. For N-UCLA which is a small-scale dataset, we also get state-of-the-art performance in Table 4. Our approach is 1.1% better than VPN [10]. At last, we visualized the weights learned by our method, as shown in Fig.5. It can be seen from this figure that the weights learned by the third and second partitions in drop trash action is relatively large, which is in line with our expectations. In addition, we can observe that weights predicted at action-related intervals are much larger than those of action-unrelated intervals (0.2586 vs 0.0005), which further proves the effectiveness and reasonability of our approach.

Table 3. Comparisons of the validation accuracy with state-of-the-art methods on the NTU RGB+D 120 dataset.

Method	Ske.	RGB	Att	CS_1	CS_2
ST-LSTM	√	×	√	55.7	57.9
Two stream Att LSTM	√	×	√	61.2	63.3
PEM	√	×	√	64.6	66.9
2s-AGCN	√	×	√	82.9	84.9
Two-streams+ST-LSTM	√	√	×	61.2	63.1
Separable STA	√	√	√	83.8	82.5
VPN	√	√	√	86.3	87.8
ours	√	√	√	91.6	91.9

Table 4. Comparisons of the validation accuracy with state-of-the-art methods on the N-UCLA.

Method	Ske.	RGB	Att	CV
Glimpse clouds	√	√	√	90.1
Separable STA	√	√	√	92.4
P-I3D	√	√	√	93.1
VPN	√	√	√	93.5
ours	√	√	√	94.9

5. CONCLUSION

In this work, we propose a novel two-stream network which consists of RGB stream and skeleton stream. Our methods put forward a special attention module which is designed to learn the weights of three partitions in different times. Through our method, the information of areas associated with the action in key interactions can be selected accurately.

6. ACKNOWLEDGEMENT

This work was supported by Key-Area Research and Development Program of Guangdong Province

(No.2021B0101410003) and National Natural Science Foundation of China under Grants 62006230, 62076235, 61976210 and U21B2043.

7. REFERENCES

[1] Q. Chen and Y. Zhang, "Sequential segment networks for action recognition," *IEEE Signal Processing Letters*, vol. 24, pp. 712–716, 2017.

[2] S. Yan, Yuanjun Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," *ArXiv*, vol. abs/1801.07455, 2018.

[3] Jiagang Zhu, Wei Zou, Zheng Zhu, Liang Xu, and Guan Huang, "Action machine: Toward person-centric action recognition in videos," *IEEE Signal Processing Letters*, vol. 26, no. 11, pp. 1633–1637, 2019.

[4] L. Shi, Yifan Zhang, Jian Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12018–12027, 2019.

[5] Ke Cheng, Yifan Zhang, X. He, Wei-Han Chen, Jian Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 180–189, 2020.

[6] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[7] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[9] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4597–4605.

[10] Ji Lin, Chuang Gan, and Song Han, "Tsm: Temporal shift module for efficient video understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7083–7093.

[11] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein, "Geometric deep learning on graphs and manifolds using mixture model cnns," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5115–5124.

[12] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, 2017, vol. 31.

[13] William L Hamilton, Rex Ying, and Jure Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[14] Amir Shahroudy, Gang Wang, and Tian-Tsong Ng, "Multi-modal feature fusion for action recognition in rgb-d sequences," in *2014 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*. IEEE, 2014, pp. 1–4.

[15] Guiyu Liu, Jiuchao Qian, Fei Wen, Xiaoguang Zhu, Rendong Ying, and Peilin Liu, "Action recognition based on 3d skeleton and rgb frame fusion," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 258–264.

[16] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 833–842.

[17] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat, "Vpn: Learning video-pose embedding for activities of daily living," in *European Conference on Computer Vision*. Springer, 2020, pp. 72–90.

[18] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 784–799.

[19] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[20] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[21] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu, "Cross-view action modeling, learning and recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2649–2656.