VIDEO-BASED EARLY ASD DETECTION VIA TEMPORAL PYRAMID NETWORKS

Yuan Tian[†], *Xiongkuo Min*[†], *Guangtao Zhai*[†], *and Zhiyong Gao*[†]

[†]Institute of Image Commu. and Network Engin., Shanghai Jiao Tong University, China
Email: {ee_tianyuan, minxiongkuo, zhaiguangtao, zhiyong.gao}@sjtu.edu.cn

ABSTRACT

Autism spectrum disorder (ASD) is a brain-based disorder characterized by social deficits and repetitive behaviors, highrising in children. In this work, we first build a ASD video dataset, and then introduce the one glimpse early ASD detection (O-GAD) network, an effective and efficient end-to-end deep architecture for video-based early ASD detection. Our network can take arbitrary-length videos as input, detecting ASD typical actions and determining if repetitive behaviours appeared only at one glimpse. The experimental results show that our method outperforms other state-of-the-art video content analysis methods on this task in terms of mAP. Moreover, we conduct extensive ablation experiments to demonstrate the effectiveness and rationality of the designed network structure.

***Index Terms*—** ASD, Autism diagnosis, Action Detection, Video Processing, Convolutional Neural Networks

1. INTRODUCTION

Autism spectrum disorders (ASD) is a disorder, which is relatively common in children, mainly characterized by social deficits and repetitive stereotyped behaviours. Early detection for ASD can help children receive timely behavioral therapy, which can improve their daily functioning, decrease symptom serverity and optimize long-term outcomes [1].

Despite early ASD detection's critical importance, no existing methods can routinely screen early ASD effectively and in low cost. However, ASD people have some common actions and repetitive behaviours [2]. With the rise of Deep Learning and Convolutional Neural Networks (CNNs), impressive progress has been made in action detection [3] and action recognition [4], inspiring us to use the modern data-driven Deep Learning based methods on this task.

We believe that several issues have to be addressed for video-based early ASD detection: (1) Scarcity of video dataset annotated with ASD atypical action types, especially with repetitive behaviour label. (2) No existing action detection method can detect ASD atypical actions and repetitive behavior simultaneously with high recall.

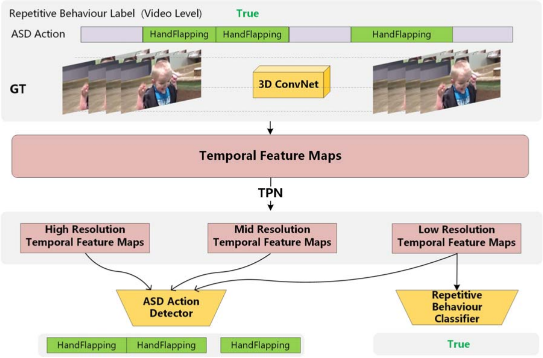
Repetitive Behaviour Label (Video Level)
True
ASD Action
HandFlapping
HandFlapping
HandFlapping
GT
3D ConvNet
Temporal Feature Maps
TPN
High Resolution Temporal Feature Maps
Mid Resolution Temporal Feature Maps
Low Resolution Temporal Feature Maps
ASD Action Detector
Repetitive Behaviour Classifier
HandFlapping
HandFlapping
HandFlapping
True

Fig. 1: Overview of our system. Given an arbitrary-length ASD surveillance video clip, we extract temporal feature maps with 3D convolutional filters, generates high-level semantic multi-resolution pyramid features by TPN 3.2, and then simultaneously predicts ASD related action and if the video clip contains high-risk stereotyped repetitive behaviour or not.

To prepare training data for this task, we build an untrimmed ASD surveillance video dataset with dense annotations. To address the second issue, we propose the one glimpse early ASD detection (O-GAD) network, which is a temporal convolutional network detecting the ASD atypical actions and classifying repetitive behaviour simultaneously from videos at one glimpse. Inspired by some state of the art action detection methods such as C3D [5], SSAD [6] and R-C3D [7], our O-GAD network directly predicts ASD actions and repetitive behaviours, as shown in Figure 1. Our network accepts video of arbitrary length (only restricted by GPU memory). The key contributions of our work are the following:

- To the best of our knowledge, our work is the first video-based early ASD diagnosis method not requiring subjects performing special actions, which can screen early ASD from suspected patient's surveillance video.
- We build an ASD video dataset with a total length of 40 hours, named $ASD40h$. We believe this dataset can eliminate the gap between appearance-based ASD diagnosis research and real applications to some extent.
- In this work, we adopt 3D CNN to generate shared temporal feature maps from videos. In subsequent network layers, we propose a novel temporal pyramid network (TPN3.2) to dig pyramid features of different semantic levels from temporal feature maps. We utilize these pyramid features for tasks of different granularity - short-term ASD related action detection and long-term repetitive behaviour recognition.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00055

2. RELATED WORK

The Importance of Early ASD Detection and Routine Screening. [8][9][10] reflect a 10-fold confirmed identification increase of ASD from studies published a half-century ago, especially in *1-6* years-old children. [11] pointed that early identification of ASDs allows early intervention, etiologic investigation, and counseling regarding recurrence risk.

Identification of ASD. Psychologists have conducted extensive studies on this field. Lord and Pickles [12] proposed a semistructured, standardized assessment, termed ADOS exam, consisting of four 30-minute modules to predict the level of social deficits. The long-length of ADOS exam (spans serval hours) and the requirements of professional psychologist checking each exam modules mean it cannot be widely used for early screening. Recently, advances in magnetic resonance imaging (MRI) and genetic engineer have helped diagnose ASD more effectively. [13] reviewed the studies on brain connectivity changes in ASD using either resting state functional MRI or diffusion tensor imaging. [14] leveraged genetic findings through advances in model systems and integrated genomic approaches to develop new classes of therapies and personalised approachs to treatment, but these diagnoses are less sensitive to early ASD and can't be utilized as a routine screening way for the high cost. [15] studied that experienced observers may be able to distinguish children with ASD only by donated family videos, inspiring us to adopt deep-learning based video content analysis methods to address the early ASD screening problem. Unlike [16], we don't require subjects performing special actions. So, our method can server as routine screening.

Video-based Action Recognition and Detection. Action recognition is an important research topic for video content analysis. We mainly review the state-of-the-art deep-learning based methods. Two-stream network [4, 17] learns both spatial and temporal features by operating network on single frame and stacked optical flow field respectively using 2D Convolutional Neural Network (CNN). C3D network [5] used 3D convolution to capture both spatial and temporal information directly from raw video. Recently, the temporal action detection task has been studies extensively, focusing on how to detect action instances in untrimmed videos where the boundaries and categories of action instances have been annotated. Recurrent Neural Network (RNN) is widely used in many action detection approaches [18, 19, 20] to encode feature sequence and make per-frame prediction of action categories. However, it is difficult for RNNs to keep a long time period memory in practice [20]. An alternative choice is temporal convolution. For example, Lea et al. [21] proposes Temporal Convolutional Networks (TCN) for temporal action segmentation. Our O-GAD network is also mainly composed of temporal convolutional layers, aiming to handle input video snippets of long-length.

3. METHODOLOGY

We propose an one glimpse early ASD detection (O-GAD) network, a novel convolutional neural network for early ASD detection in surveillance video streams. The network, illustrated in Figure 2, consists of four components: a 3D ConvNet temporal feature extractor, a temporal pyramid network, an ASD action detector and a repetitive behaviour discriminator. To enable efficient computation and end-to-end training, the two tasks share the same C3D temporal feature maps. The temporal pyramid network use temporal convolution to continually shorten the feature map and output anchor feature maps for subsequent detection tasks, while keeping the high-level semantics of shadow layer. The ASD action detector classifies the anchor segments into multiple categories and refines the segment boundaries in one stage. The repetitive behaviour discriminator discriminates whether the whole video contains repetitive behaviour.

3.1. Temporal Feature Extractor

Video Temporal Feature Extractor can be divided into two parts, the 3D spatial-temporal feature extractor and the subsequent temporal feature extractor.

It has been shown that both spatial and temporal features are important for representing videos, and a *3D ConvNet* encodes both spatial and temporal features in a convolutional fashion. We use a *3D ConvNet* to extract rich spatio-temporal feature from a given input video buffer, i.e., a sequence of RGB video frames with dimension $\mathbb{R}^{3\times L\times 112\times 112}$. The architecture of the *3D ConvNet* is taken from the C3D architecture [5]. We adopt the convolutional layers (`conv1a` to `conv5b`) of C3D, so a feature map $C_{conv5b} \in \mathbb{R}^{512\times \frac{L}{8}\times 7\times 7}$ is produced.

To obtain *temporal* only features, like in [7], we first add a 3D convolutional filter with kernel size $3 \times 3 \times 3$ with stride 2 on top of C_{conv5b} to extend the temporal receptive field. Then, we apply a 3D max-pooling filter with kernel size $1 \times 7 \times 7$ to downsample the feature's spatial dimensions from 7×7 to 1×1. We denote the final temporal feature as f_{TC0}.

3.2. Temporal Pyramid Network

After extracting temporal only features from input video, we stack two anchor convolutional layers (*Temporal Conv1 (TConv1)* and *Temporal Conv2 (TConv2)*) on them. These layers have same configuration: kernel size 3, stride size 2 and 512 convolutional filters. The output feature maps of anchor layers are f_{TC1} and f_{TC2} with size $(T_w/32 \times 512)$ and $(T_w/64 \times 512)$ respectively. f_{TC0}, f_{TC1} and f_{TC2} are in a pyramidal feature hierarchical fashion. However, the *ConvNets* usually have semantics from low to high levels. Namely, the semantics progressively weaken from f_{TC0} to f_{TC2}, leading to the precision loss in ASD actions predicted from f_{TC1} and f_{TC2}. Inspired by FPN [22], we develop a

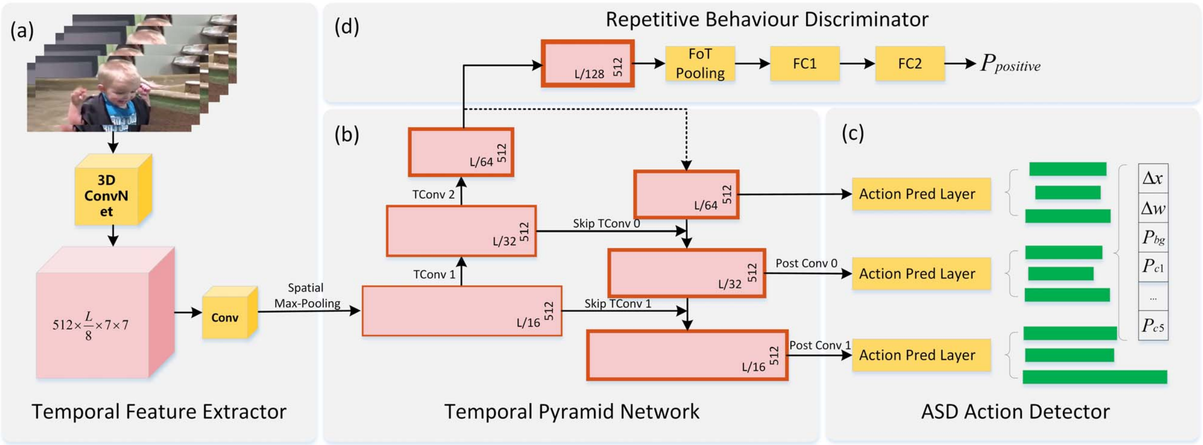

Fig. 2: The framework of our approach. (a) 3DConv and spatial max-pooling are used to extract temporal feature. Temporal features are denoted as pink blocks. (b) TPN is developed to build high-level semantics temporal feature maps at all scales. Thicker outlines of feature maps denote semantically stronger feature. (c) ASD Action Detector predicts action snippets (boundaries and action category scores) from different level of FPN output temporal feature maps. (d) Repetitive Behaviour Discriminator discriminates the input video of high-risk repetitive behaviour.

full-convolutional temporal pyramid network (TPN) with skip connections to build high-level semantics temporal feature maps at all scales, with marginal extra cost of computation and memory. TPN consists of top-down pathway, bottom-up pathway and the skip connections between two pathways. The bottom-up pathway is the feed-forward computation of anchor layers. The top-down pathway hallucinates higher resolution features by upsampling (with a factor of 2) temporally coarser, but semantically stronger, feature maps from lower levels. These features are then enhanced with features from the bottom-up pathway via skip connections. We adopt temporal convolutional filters with kernel size 3, stride size 1 to build the skip connection, preventing the top-down path's gradients from propagating to feed-forward conv layers directly. The upsampled map is then add to the corresponding feature maps produced by skip convolutional filters in a element-wise fashion. Finally, we append a temporal convolution with kernel size 3 on each merged map to generate the final feature map, which is to reduce the aliasing effect of upsampling. We denote the final anchor feature maps as f_{TP0}, f_{TP1} and f_{TP2}.

3.3. ASD Action Detector

For each temporal feature map of anchor layers, we associate a set of multiple scale ASD action anchor instances with each feature map cell. Considering the anchor feature map's smaller resolution and larger receptive field, we let lower anchor layers ($TP1$ and $TP2$) to predict longer ASD action (e.g. *moving fingers in front of the eyes*) instances while top anchor layer($TP0$) to predict short ASD action (e.g. *hand flapping*) instances.

Considering an anchor feature map with length W_f, we define the base scale $s_f = \frac{1}{W_f}$, and a set of temporal scale ratios $R_f = \{r_d\}_{d=1}^{D_f}$, where D_f is the number of scale ratios. Then the number of associated anchor instances with the an anchor feature map is $W_f \cdot D_f$. For each ratio r_d, we calculate $\mu_w = s_f \cdot r_d$ as anchor instance's default width. Nonetheless, all anchor instances associated with the i-th feature map cell shared the same default center location $\mu_c = ceil(\frac{i}{W_f})$. Empirically, we let $R_f \in \{0.5, 0.75, 1, 1.5, 2\}$.

The anchors tile the feature map in a convolutional manner, so that the position of each anchor relative to its corresponding cell is fixed. We adopt a set of convolutional filters with kernel size 3, stride size 1 for predicting action category scores and two location offsets relative to each anchor instance. This results in a total of $(k+2)\cdot D_f$ filters that are applied around each cell in the feature map, yielding $(k+2)\cdot D_f \cdot W_f$ ouputs for a W_f length temporal feature map. For ASD action detector, we take 5 typical ASD action in consideration, meaning $k = 6$. We denote the prediction vector of each anchor instances as $\boldsymbol{p_{pred}} = (\boldsymbol{p_{class}}, \Delta x, \Delta w)$ with length $k + 2$. $\boldsymbol{p_{class}}$ is classification score used to predict anchor instance's category. Δx, Δw are location offsets used for adjusting the default location of anchor instance. The adjusted location is defined as:

$$\begin{aligned} x &= \mu_x + \alpha_1 \cdot \mu_w \cdot \Delta x \\ w &= \mu_w \cdot exp(\alpha_2 \cdot \Delta w), \end{aligned} \tag{1}$$

where x and w are center location and width of anchor instance respectively. α_1 and α_2 are used for controlling the effect of location offsets to make prediction stable. We set both α_1 and α_2 to 0.1. The starting and ending time of action instance are $x_1 = x - \frac{1}{2}\cdot w$ and $x_2 = x + \frac{1}{2}\cdot w$ respectively. So for a anchor feature map f, we can get a ASD action instances set $\Phi_f = \{\phi_n = (x, w, \boldsymbol{p_{class}})\}_{n=1}^{N_f}$, where $N_f = D_f \cdot W_f$ is the number of anchor instances. The overall prediction of ASD actions is $\Phi_A = \{\Phi_{f_{TP0}}, \Phi_{f_{TP0}}, \Phi_{f_{TP0}}\}$.

3.4. Repetitive Behaviour Discriminator

Since the video clip input to our network is about $1024frames \approx 41s$ and the length of the annotated repetitive behaviours vary from $18.2s$ to $50s$, we treat the discrimination of repetitive behaviours as a classification problem.

We have temporal feature maps f_{TC3} of high-level semantics. However, we need to extract fixed-length features for each of them in order to use fully connected layers for further repetitive behaviour classification task. Inspired by [23] and [7], we design a fixed-length output temporal pooling layer, termed *FoT pooling*, to extract the fixed-length temporal features from f_{TC3}. Specifically, in FoT pooling, the input temporal feature of length l is divided into 6 sub-segments, each with approximate size $\frac{l}{6}$, then max pooling is performed inside each temporal segment. Thus, arbitrary-length videos give rise to temporal features of the same size $512 \times 6 \times 1 \times 1$ after performing FoT pooling. The output of the FoT pooling is then fed to two fully connected layers. Here, the video clip are classified to whether containing repetitive behaviour by a classification layer .

3.5. Network Training

Data Preparation. Firstly, restricted by the GPU memory, we first split every video into 1024 frame snippets with 80% overlap. In terms of the video snippet s.

For the task of detecting ASD actions, ASD action instances set Φ_A is computed via ASD action detector network during training. We need to match them with ground truth set Φ_s for label assignment. For an anchor instance ϕ_i in Φ_A, we calculate it's IoU overlap with all ground truth instances in Φ_s. If the highest IoU overlap is higher than 0.5, we match $\phi_i = (\boldsymbol{p_{class}}, \Delta x, \Delta w)$ with the corresponding ground truth instance $\phi_g = (c_g, g_x, g_w)$ and regard it as positive, otherwise negative, where c_g is the category of ϕ_g and is set to 0 for negative instance, g_x and g_w are center location and width of ϕ_g respectively. For the task of recognizing the repetitive behaviour, if the input video contains more then half of one repetitive behaviour, we assign it with positive label, otherwise negative.We select the negative samples randomly to keep the ratio between positive and negative instances be nearly 1:1 for above tasks.

Training Objective. The training objective of the O-GAD network is to solve a multi-task optimization problem. The overall loss function is the weighted sum of three parts: (1) the ASD related action classification loss (action class); (2) the ASD related action detection loss (action loc); (3) and the repetitive behaviour probability loss (repetitive class). The first two are for the ASD related action detection task and the last one is for the repetitive behaviour classification task. We denote the loss function as:

$$L = L_{action\ class} + \alpha \cdot L_{action\ loc} + \beta \cdot L_{repetitive\ class}, \quad (2)$$

where α and β are the weight terms used for balancing each part of loss function. Empirically,we set α to 10 and β to 1 by cross validation. Let $x_{ij}^p = \{1, 0\}$ be an indicator for matching the i-th default ASD related action anchor to the j-th ground truth action of category k. For the action classification loss, we use conventional softmax loss over multiple categories, which is effective for training classification model and can be defined as:

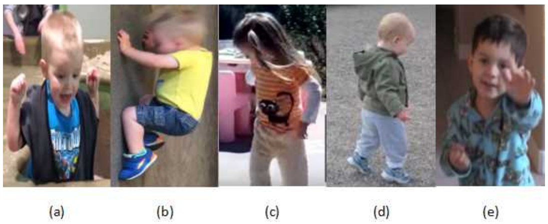

Fig. 3: ASD action types. (a) Hand Flapping. (b) Head Banging. (c) Spinning in a Circle. (d) Toe Walking. (e) Moving Fingers.

$$L_{action\ class} = -\frac{1}{N}\sum_{i=1}^{N} x_{ij}^{p} log(P_i^k). \quad (3)$$

where $P_i^{(k)} = \frac{exp(p_i^{(k)})}{\sum_k exp(p_i^{(k)})}$. N is the number of matched default temporal anchors, If $N = 0$, we set the $L_{action\ class}$ to 0. $L_{action\ loc}$ is the Smooth L1 loss [23] for location offsets. We regress the center (x^*) and width (w^*) of predicted instance:

$$L_{action\ loc} = \frac{1}{N}\sum_{i \in Pos}^{N} (smoothL_1(x_i^* - x_i) + smoothL_1(w_i^* - w_i)). \quad (4)$$

where x_i and w_i is the center location and width of ground truth instance. For repetitive behaviour classification loss, we implement it as a cross entropy loss for simplicity.

4. EXPERIMENTS

We empirically evaluate the effectiveness of our method on ASD early detection task. The experimental settings and results are described in this section.

Dataset. We build an untrimmed video dataset of 30 videos (about 40h in whole) named $ASD40h$, containing 5 most common ASD atypical action classes, hand flapping, head banging, spinning in a circle, toe walking and moving fingers in front of the eyes, shown in Figure 3. Each video was annotated with both atypical action instances and repetitive behaviour instances. The video dataset was divided into 20 validation and 10 test video sequences.

Data Augmentation. Firstly, we utilize the state of the art segmentation network-Deep Lab v3+ [24] to segment the person instance from the image and then overlay these foreground regions on some scene backgrounds to increase the diversity of the backgound. Secondly, we use shear transformations with shear parameter $\theta \in [-25°, 25°]$ to efficiently simulate limited viewpoint changes. We conduct the above processing methods on every frame of video. Besides, we also leverage random flipping and corner cropping like [5][7].

Implementation Details. For training of the O-GAD network, we use the Adam algorithm [25] with the aforementioned multi-task loss function. Our implementation is based on PyTorch [26], with training executed on GeForce 1080Ti GPUs. For the initialization of temporal-spatial network, we adopt the 3D CNN parts of the pre-trained C3D [5] model on sport1m[27]. We adopt the Xavier method [28] to randomly initialize temporal networks.

4.1. Comparison with other State-of-the-arts

We compare our method with other state-of-the-art video content analysis methods on $ASD40h$. Note the fact that (1) the length of the input video clips are 41 seconds. (2) The average ASD action duration in $ASD40h$ is 3.6 seconds. (3) The average annotated repetitive behaviour duration is 28.4 seconds. Thus, the task of video-based ASD screening contains two parts, the ASD short-length action detection task and the long-length repetitive behaviour recognition task.

Action detection's perspective. We report the ASD action detection performances measured by mean average precision (mAP) for different IoU thresholds α. We measure S-CNN [29], SST [18]+S-CNN classifier, R-C3D [7] and SSAD [6] on $ASD40h$.

The comparison results between our O-GAD network and other state-of-the-art systems are shown in Table 1 with multiple overlap IoU thresholds varied from 0.1 to 0.5. We can find from the results that our method's mAP@0.5 surpass the most state-of-the-art two-stage action detector R-C3D by almost **1%**. For the reason that R-C3D designs the proposal subnet explicitly, the recall value of the R-C3D's proposal module is 86% compared to 79% on our method. This introduces our method's lower mAP@0.1.

We also treat repetitive behaviour as a special action category (only this category and background in this experiment), and set the IoU threshold to 0.2. The evaluation results on repetitive behaviour are shown in Table 2.

Action Recognition's perspective. As for repetitive behaviour recognition task, we compare O-GAD's recognition branch with a few baselines: the current best hand-crafted feature iDT [30], two-stream network [4], and the spatial-temporal C3D network [5]. Unlike other state-of-the-art action recognition methods, which predict label scores from each short-term snippets (16 Frames for C3D) and fusion the scores to final label, our method directly predicts the final score. The comparison results are shown in Table 3. Our method outperforms C3D over **1%**. We train the C3D network improved with *FoT Pooling* to accept long-length video, further validating O-GAN's superiority (**0.7%** improvement). We also trained our O-GAN network's recognition branch separately, showing a 0.3% performance decrease compared to full O-GAD network. We believe the multi-task learning strategy contributes to the performance gap.

4.2. Ablation Experiments

We conduct ablative experiments to understand the effects of the different components and parameters in our framework, including temporal feature extractor and architectures of TPN network. In this section, We utilize the mAP@0.5 and mAP as the metric of ASD action detection and repetitive behaviour recognition task respectively.

Table 1: mAP results on $ASD40h$(repetitive behaviour excluded) with various IoU threshold α used in evaluation.

α	0.1	0.2	0.3	0.4	0.5
S-CNN[29]	49.3	45.3	37.8	30.1	21.8
SST[18][1]	50.4	47.4	44.3	32.7	25.5
SSAD[6]	53.0	50.4	47.8	34.5	25.7
R-C3D[7]	**58.3**	53.6	48.2	34.8	27.2
Proposed	56.6	**54.2**	**50.1**	**35.2**	**28.1**

Table 2: mAP results on ASD40h repetitive behaviour recognition result.

Method	S-CNN[29]	SST[18]	SSAD[6]	R-C3D[7]	Proposed
mAP	60.2	62.3	66.5	65.3	**70.8**

Table 3: Performance of repetitive behaviour recognition task on ASD40h among different methods.

Method	Accuracy(%)
(iDT)[30]	88.2
two-stream network [4]	89.0
C3D network[5](16 Frames+laterly Fusion)	93.6
C3D network[5](FoT Pooling)	94.5
Proposed(recognition branch)	94.9
Proposed	**95.2**

Table 4: Performance impact from several temporal feature extraction methods.

Method	3D Conv	Max Pooling	Average Pooling
mAP@0.5	**28.1**	23.9	23.6
mAP	94.6	**94.9**	94.8

Table 5: Evaluation of ASD action detection on different TPN settings.

Base TPN		✓	✓	✓
Skip Conv		✓		✓
Post Conv			✓	✓
mAP@0.5	24.1	27.0	27.2	**28.1**

Temporal Feature Extractor. For extending the temporal receptive field further on top of the C3D feature, we tried three post temporal fusion methods, $3 \times 3 \times 3$ 3D convolutional filters with stride size 2, $2 \times 1 \times 1$ temporal max pooling and $2 \times 1 \times 1$ temporal average pooling. The comparison results are shown in Table 4. We can find that 3D Conv has obvious advantages than pooling, 0.2% decrease on recognition task, but over **4%** performance boost on detection task! The experiment results is reasonable: (1) our C3D base network is initialized from the pre-trained model targeted for sport1m[27] action classification task, which has semantic deviation from action detection task. The 3D Conv layer acts as a transfer learning layer. (2) Video data includes redundant information. The max pooling layer can select more discriminative features and therefore has slight performance advantage over 3D Conv and average pooling.

TPN Network Architecture. In this experiment, we proved TPN's contribution to the improvements of final results and evaluated multiple variants of TPN. The results are shown in Table 5. We can find that single using skip connection conv layers has better performance than single using anti-aliasing post conv layers. The O-GAD network achieves the best performance with the full TPN including both two types of conv layers.

5. CONCLUSION

In this paper, we build the first ASD video dataset with dense annotations and propose the one glimpse early ASD detection(O-GAD) network, the first end-to-end video-based early ASD detection architecture. Our O-GAD network directly predicts ASD actions and repetitive behaviours from surveillance videos. Our architecture outperforms other state-of-the-art methods by **0.9%** and **1.6%** on two tasks respectively. One future direction may be trying to distinct ASD people from normal people directly from video, bringing more powerful guidance to ASD diagnosis.

6. REFERENCES

[1] Mona Al-Qabandi, Jan Willem Gorter, and Peter Rosenbaum, "Early autism detection: are we ready for routine screening?," *Pediatrics*, pp. peds–2010, 2011.

[2] Christian R Marshall, Abdul Noor, John B Vincent, Anath C Lionel, Lars Feuk, Jennifer Skaug, Mary Shago, Rainald Moessner, Dalila Pinto, Yan Ren, et al., "Structural variation of chromosomes in autism spectrum disorder," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 477–488, 2008.

[3] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2678–2687.

[4] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[5] 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.

[6] Tianwei Lin, Xu Zhao, and Zheng Shou, "Single shot temporal action detection," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 988–996.

[7] Huijuan Xu, Abir Das, and Kate Saenko, "R-c3d: region convolutional 3d network for temporal activity detection," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 5794–5803.

[8] Gillian Baird, Tony Charman, Simon Baron-Cohen, Antony Cox, John Swettenham, Sally Wheelwright, and Auriol Drew, "A screening instrument for autism at 18 months of age: a 6-year follow-up study," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 39, no. 6, pp. 694–702, 2000.

[9] Suniti Chakrabarti and Eric Fombonne, "Pervasive developmental disorders in preschool children," *Jama*, vol. 285, no. 24, pp. 3093–3099, 2001.

[10] Suniti Chakrabarti and Eric Fombonne, "Pervasive developmental disorders in preschool children: confirmation of high prevalence," *American Journal of Psychiatry*, vol. 162, no. 6, pp. 1133–1141, 2005.

[11] Chris Plauché Johnson, Scott M Myers, et al., "Identification and evaluation of children with autism spectrum disorders," *Pediatrics*, vol. 120, no. 5, pp. 1183–1215, 2007.

[12] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H Cook, Bennett L Leventhal, Pamela C DiLavore, Andrew Pickles, and Michael Rutter, "The autism diagnostic observation schedulegeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[13] Pallavi Rane, David Cochran, Steven M Hodge, Christian Haselgrove, David Kennedy, and Jean A Frazier, "Connectivity in autism: a review of mri connectivity studies," *Harvard review of psychiatry*, vol. 23, no. 4, pp. 223, 2015.

[14] Daniel H Geschwind and Matthew W State, "Gene hunting in autism spectrum disorder: on the path to precision medicine," *The Lancet Neurology*, vol. 14, no. 11, pp. 1109–1120, 2015.

[15] Alison Kerr, "Annotation: Rett syndrome: recent progress and implications for research and clinical practice," *Journal of Child Psychology and Psychiatry*, vol. 43, no. 3, pp. 277–287, 2002.

[16] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, Jessica Podda, Francesca Battaglia, Edvige Veneselli, Cristina Becchio, and Vittorio Murine, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3421–3426.

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao, "Towards good practices for very deep two-stream convnets," *arXiv preprint arXiv:1507.02159*, 2015.

[18] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles, "Sst: Single-stream temporal action proposals," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6373–6382.

[19] Shugao Ma, Leonid Sigal, and Stan Sclaroff, "Learning activity progression in lstms for activity detection and early detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1942–1950.

[20] Bharat Singh, Tim K Marks, Michael Jones, Oncel Tuzel, and Ming Shao, "A multi-stream bi-directional recurrent neural network for fine-grained action detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1961–1970.

[21] Colin Lea, Rene Vidal, Austin Reiter, and Gregory D Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[22] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie, "Feature pyramid networks for object detection.," in *CVPR*, 2017, vol. 1, p. 4.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[24] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[27] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[28] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[29] Zheng Shou, Dongang Wang, and Shih-Fu Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1049–1058.

[30] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3551–3558.