Where are the Children with Autism Looking in Reality?

Xiaoyu Ren[1], Huiyu Duan[1], Xiongkuo Min[1], Yucheng Zhu[1], Wei Shen[1], Linlin Wang[2], Fangyu Shi[1], Lei Fan[1], Xiaokang Yang[1], and Guangtao Zhai[1(✉)]

[1] Shanghai Jiao Tong University, Shanghai, China
{windkaiser,huiyuduan,minxiongkuo,zyc420,wei.shen,fangyu.shi,lei.fan,xkyang,zhaiguangtao}@sjtu.edu.cn
[2] Shanghai Donglifengmei School, Shanghai, China

Abstract. Social difficulties are hallmarks of individuals with autism spectrum disorder (ASD), of which atypical visual attention is one of the most important characteristics. Learning and modeling the atypical visual attention of individuals with ASD have particularly important significance to related research in the fields of medical science, psychology, education *etc.*, and many studies have been conducted in the literature. However, previous studies have two weaknesses. First, all stimuli in the conducted experiments are selected by the researchers, which are not only restricted by the objective and intention of the researchers, but also limited by the subjective cognition of the photographers. Secondly, most of these stimuli are displayed on screens with restricted and relatively small field-of-view (FOV) compared with the real world. Therefore, in this paper, we conduct the first large-scale study towards better understanding and modeling the atypical visual attention of individuals with ASD in real world. To overcome the two weaknesses mentioned above, a large-scale dataset is established which includes 300 omnidirectional images with the corresponding eye tracking data collected under virtual reality (VR) environment among 15 children with ASD and 16 typically developing (TD) controls. Moreover, a vector quantized saliency prediction model (VQSAL) is applied to better learn the visual attention patterns of both ASD and TD people under the omnidirectional condition.

Keywords: Autism spectrum disorder (ASD) · Atypical visual attention · Virtual reality (VR)

1 Introduction

Autism is a complex neurodevelopmental condition, and little is known about its neurobiology [1]. Phenotype markers including social communication symp-

This work was supported by National Key R&D Program of China 2021YFE0206700, NSFC 61831015, 61901260, 62176159, 62101326, Natural Science Foundation of Shanghai 21ZR1432200, Shanghai Municipal Science and Technology Major Project 2021SHZDZX0102, China Postdoctoral Science Foundation 2022M712090.

X. Ren and H. Duan—Equal contribution.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 588–600, 2022.
https://doi.org/10.1007/978-3-031-20500-2_48

toms, fixated or restricted behaviors or interests, hyper- or hypo- sensitivity to sensory stimuli, and associated features have been widely used in characterizing and diagnosing the Autism Spectrum Disorder (ASD) [2]. Among these, social difficulties are known as the hallmark features of autism. As an important aspect of social difficulties, atypical visual attention is frequently observed in individuals with autism [3] and reported in the literature [4,5]. Several possibly related visual attention traits of autistic individuals have been reported in some early studies, including reduced joint-attention behaviours [6], reduced attention to social stimuli (*i.e.*, faces, conversations, *etc.*) but increased attention to non-social stimuli (*i.e.*, vehicles, electronics, *etc.*) [7,8], reduced visual attention to core facial features [9,10], *etc.* However, the vast majority of these prior studies have used restricted or unnatural stimuli, which limited the exploration of the common characteristics underlying the ASD.

Recently, some studies have conducted large-scale experiments for characterizing the visual attention traits of ASD. Wang *et al.* [4] have quantified the atypical visual attention in ASD across multi-level features using natural stimuli and pointed out the preference of individuals with autism to low-level features of the stimuli. Jiang *et al.* [11] have presented a method to model the visual attention differences between individual with ASD and healthy people. Duan *et al.* [12] have established a large-scale open eye movement dataset for children with autism and fine-tuned four state-of-the-art (SOTA) visual attention models for learning the gaze pattern of autistic children [13]. Duan *et al.* [14] have further conducted a large-scale eye movement study for children with autism on face stimuli and proposed a model to characterize gaze pattern under this specific condition. Fang *et al.* [15] have studied the visual attention of children with autism on gaze-following stimuli and proposed a LSTM-based saliency model for classifying the gaze patterns between autistic children and typically developing (TD) controls. However, though these studies have conducted large-scale experiments on natural stimuli, all these stimuli were limited by the intended selections of the researchers and the restricted fields-of-view (FOVs) from the photographers, and neither of them has autism. The omnidirectional visual attention characteristics of ASD are still unknown. Furthermore, all these stimuli were displayed on the relatively small screens, while the differences between the semantic-level perception of screen images and real world still exist.

These two weakness of previous studies motivate us to conduct this study, *i.e.*, understanding and modeling the gaze pattern of children with autism in the real world. Instead of displaying image stimuli on screens, the eye tracking experiments in this work are conducted in Virtual Reality (VR) environment. A large-scale eye movement dataset is first established, which includes 300 omnidirectional images with the corresponding eye movement data collected from 15 children with autism and 16 TD controls. Based on the dataset, we further analyze the gaze pattern differences between autistic children and healthy children under this nearly natural condition. Moreover, we also apply a saliency prediction model based on the vector quantized neural network for better modeling the visual attention of both ASD and TD people under omnidirectional condition.

To the best of our knowledge, this is the first study that analyzes and models the omnidirectional visual attention of children with autism in the literature towards better understanding the gaze pattern of them. Eye movements encode rich information about the attention, cognition and psychological factors of an individual. Thus, understanding and modeling the gaze pattern of children with autism in (virtual) reality can not only help to further understand autism, but also may contribute to related application areas, such as diagnosis [11,15] and rehabilitation [16].

2 Subjective Experiment and Analysis

2.1 Omnidirectional Image Stimuli and Displaying Apparatus

We collected 300 omnidirectional images with high-resolution from two large-scale 360 image databases, including 85 images from Salient360 [17] and 215 images from SUN360 [18]. As shown in Fig. 1, the collected images contain various scenes in indoor and outdoor scenarios. Moreover, considering the differences between the visual attention of individuals with/without autism to social/non-social stimuli, we also balanced the semantic information inside the omnidirectional images. As shown in Fig. 1, our stimuli include rich visual features with various pixel-level, object-level, and semantic-level information.

We used HTC VIVE Pro Eye[1] as the hardware apparatus to display omnidirectional stimuli and collect eye movement data [19–23]. The software system was designed using Unity3D[2] to control the experimental procedure and record all data. The resolution of the display inside HTC VIVE Pro Eye is 1440×1600 pixels per eye which covers 110° FOV. The refresh rate 90 Hz. The eye-tracker inside it is supported by Tobii, and the frequency to collect gaze data 120 Hz.

2.2 Subjects

We recruited 31 subjects in our experiments, including 15 children with autism and 16 TD controls. All 15 autistic children were with medium-/high- function and could cooperate with us for the experiment. The age of the participants with ASD ranged from 7 years old to 13 years old with the average age of 10.4 years old. Sixteen healthy children were recruited as controls, whose ages were ranged from 7 years old to 9.6 years old with the average of 8 years old. Besides the age, the gender, handedness, and performance IQ were also matched between two groups. Before participating in the test, the parents of subjects read and signed a consent form which explained the human study. All participants had normal or correct-to-normal visual acuity during the experiment.

[1] https://www.vive.com/us/product/vive-pro-eye/overview/.
[2] https://unity.com/.

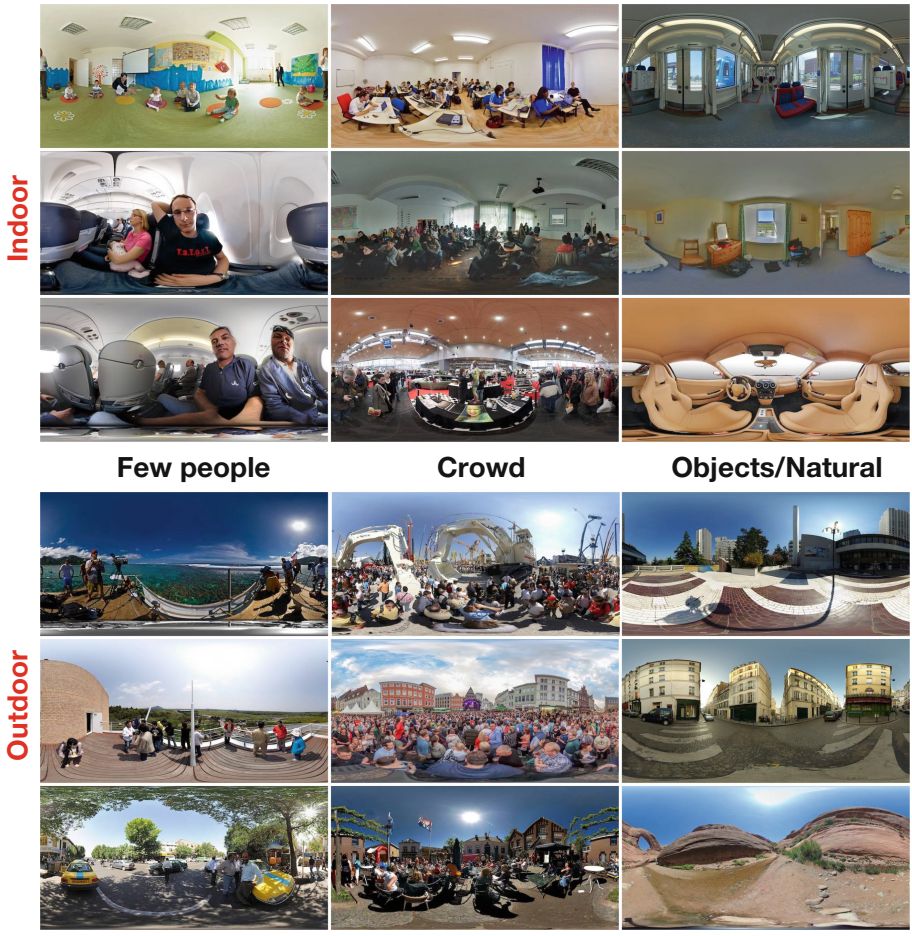

Fig. 1. Sample stimuli in our database.

2.3 Experiments

Since our work is the first study that conducts eye tracking experiments under VR environment to analyze the visual attention differences between individuals with autism and TD controls. The experiments need to be carefully designed and conducted. There were three methods in the literature to conduct eye tracking experiments under VR environment. Rai *et al.* [17] conducted the eye tracking study under seated condition with free viewing. Sitzmann *et al.* [24] studied both the seated condition and standing condition with free viewing. Haskins *et al.* [25] carried out the study under seated condition while the omnidirectional image rotate at a constant speed. Considering the possible cognition and communication problems for children with autism, in this paper we conducted the experiments in two conditions.

Standing Case. We conducted the first experiment with 200 images under standing condition with free viewing, since it is hard to teach the children with autism to use swivel chairs to look through the whole image. Due to the lack of patience of ASD participants, we split the experiment into 20 recording sessions with 10 images in each session. The initial viewing direction was initialized at the center of the omnidirectional image. Each omnidirectional image was displayed in the VR device for 20 s and followed by a 1-second gray screen mask. At the beginning of each session, we re-calibrate the eye-tracker to ensure the reliability of the acquired data.

Seated Case. Since the method in [25] may cause strong motion sickness, here we propose another way to conduct the second experiment, which includes the rest 100 images. The same as the standing case, we split the experiment into 10 recording sessions with 10 images in each session. The children were seated on a fixed chair. The initial viewing direction was similarly initialized at the center of the omnidirectional image. Each omnidirectional images was similarly displayed in the VR device for 20 s but rotated 90° every 5 s. Other procedures were the same with the standing case.

2.4 Analysis

Based on the constructed database, in this section, we analyze the differences and similarities between the visual attention of autistic children and healthy controls.

Global Comparison. We first analyze the global visual attention differences and similarities over the omnidirectional space between two groups. Figure 2 demonstrates several examples under the standing case. As shown in Fig. 2(a), children with autism try to avoid the close people who are looking at them while concentrating more on other salient targets in the car. As shown in Fig. 2(b), children with autism show reduced attention to the joint-attention of the main character in the scene but spread more attention to meaningless areas. Figure 2(c), (d) and (e) indicate the core characteristic of individuals with ASD, *i.e.*, social deficits or reduced social attention. It can be observed that children with ASD try to avoid the social targets that TD groups mainly concern and tend to concentrate on marginal directions and areas which are far way from them. It is interesting that in most of the classroom conditions, as shown in Fig. 2(h), (i), (j), the global visual attentions over the whole space are similar between the ASD group and TD group. We suppose that the semantic distributions of the classroom scenes are relatively uniform over the space, while in the scenes in Fig. 2(c), (d) and (e), the semantic distributions are spatially non-uniform. This phenomenon may reveal that the atypical visual attentions of children with autism are conditionally dependent on the scene. Moreover, autistic children and TD children show similar visual attention to non-social scenes, as shown in Fig. 2(f) and (g).

We further analyze the glocal differences over the whole space under seated condition, which are demonstrated in Fig. 3. As illustrated in Fig. 3(a) and (b), children show increased visual attention to non-social objects or information, while TD children tend to pay more attention on social information. As shown in Fig. 3, similar to the standing case, children tend to look more at the directions

Fig. 2. Global differences and similarities over the whole space between the visual attention of autistic children and healthy controls (standing case).

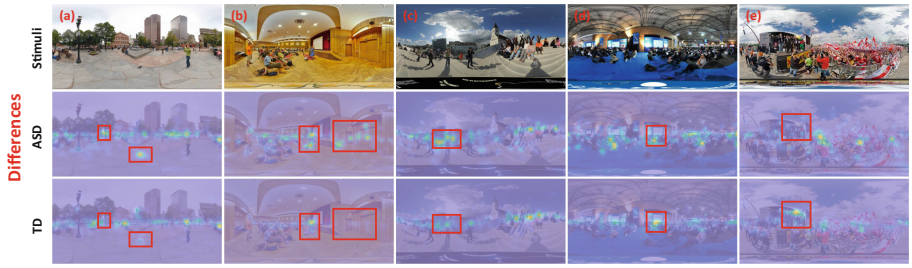

Fig. 3. Global differences over the whole space between the visual attention of autistic children and healthy controls (seated case).

far away from the social information or just trying to avoid the social scenarios near them. Moreover, children with autism seem to lack the ability of information integration and prediction, thus show reduced attention to the normally main focus in the scene as shown in Fig. 3(d) and (e).

Local Comparison. We have analyzed the global similarities and differences as above. However, we should notice that an omnidirectional image contains huge information, and even for a local FOV, the amount of information is similar to that of a regular image used in previous studies [12,13,26]. Therefore, it is also valuable to discuss the local differences between the omnidirectional visual attentions of ASD group and TD group. We show two examples in Fig. 4 to demonstrate this point. As can be observed in Fig. 4(a), in the local FOV of the yellow rectangular, children with autism tend to look more at the steering wheel,

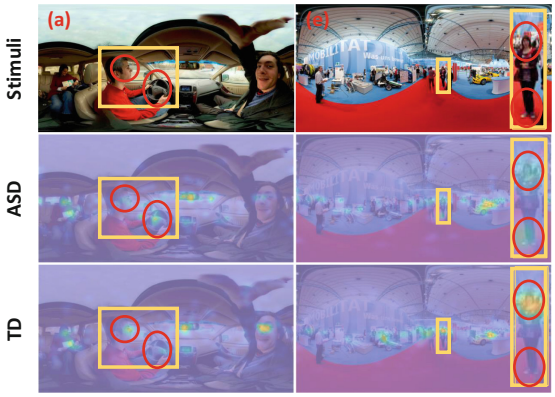

Fig. 4. Examples of local differences between the visual attention of autistic children and healthy controls for two images showed in Fig. 2 (standing case). (Color figure online)

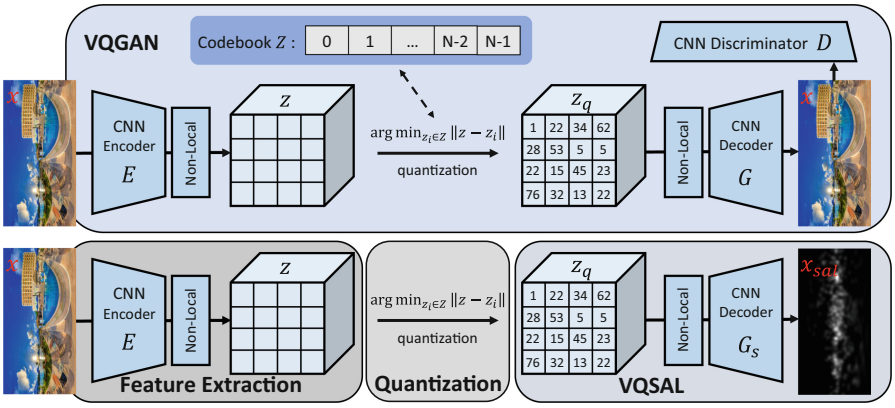

Fig. 5. Overview of the VQSAL model. Our approach uses a convolutional VQGAN to learn a context-rich codebook of the omnidirectional images, whose knowledge is then transfer to the saliency prediction.

while TD children focus more on the driver. As illustrated in the local FOV (the yellow rectangular) in Fig. 4(b), children with autism show close attention to the head and feet of the people, while TD children pay more attention to the face.

3 Omnidirectional Saliency Prediction

Our goal is to understand the context information of an image and model the visual saliency of different groups under the omnidirectional condition. In this paper, we apply a two-stage method for visual saliency prediction, which is a transfer learning framework based on a learned discrete representation model

via VQGAN, as described in Sect. 3.1. We surprisingly find this approach, summarized in Fig. 5, is harmoniously consistent with the human vision model and may be useful and reasonable for modeling the visual attention of human groups with different cognitive conditions, as discussed in Sect. 3.2.

3.1 Transfer Learning for the Saliency Prediction

We first follow the VQGAN [27] to learn a discrete representation model for omnidirectional images. Through the discrete representation model learned, we can represent any image $x \in \mathbb{R}^{H\times W\times 3}$ using a spatial collection of codebook entries $z_{\mathbf{q}} \in \mathbb{R}^{h\times w\times n_z}$ from the codebook $\mathcal{Z}$, where n_z is the dimensionality of codes and $\mathcal{Z} = \{z_k\}_{k=1}^{K} \subset \mathbb{R}_z^n$ is the learned perceptually rich code book. The representation $z_{\mathbf{q}}$ includes the extremely compressed but perceptually rich information of the image, which can be directly used to decode and predict visual saliency information. Since visual saliency is not only related to local information, but also related to global relationship, we use the non-local neural network in VQGAN to appropriately learn the global relationship and the CNN decoder to decode local information. Moreover, we use transfer learning to the decoder of the VQGAN to decode and predict the saliency density map. The overall process of this method can be represented as:

$$\hat{x}_{\text{sal}} = G_{\text{S}}(z_{\mathbf{q}}) = G_{\text{S}}\left(\mathbf{q}(E(x))\right), \tag{1}$$

where G_{S} is the decoder for saliency density prediction. The loss function of the saliency prediction in our paper is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda\mathcal{L}_{\text{sal}}, \tag{2}$$

where $\mathcal{L}_{\text{sal}} = \mathcal{L}_{\text{CC}} + \mathcal{L}_{\text{KL}}$, CC and KL are two widely used metrics for measuring the accuracy of the predicted saliency maps. The weighting factor λ is empirically set as 0.2 in this paper.

3.2 Discussion of VQSAL

We surprisingly find that this model is harmoniously consistent with the human visual attention model and may be useful and reasonable for modeling the visual attention of human groups with different cognitive conditions. The process of neural discrete inference is similar to the process of human attention formation, *i.e.*, encoding a given image to information, then quantifying the information with human knowledge base, and finally decoding to human visual attention activities. However, most of previous learnable saliency prediction methods have used only encoder [28] or autoencoder [29] to model the visual attention. Moreover, it seems unreasonable to finetune the whole model for human groups with different cognitive conditions since their vision perception systems are similar (*i.e.*, encoder part) [13]. This is precisely the rationality of our method, since the discrete representation encoder is pre-trained, and it is reasonable to simulate the visual saliency of humans with different cognitive conditions using different decoders.

Table 1. Quantitative comparison results of different models for saliency prediction on ASD group (Learnable models are fine-tuned on our dataset).

Metric\Model	Itti	GBVS [30]	GBVS360 [31]	BMS [32]	BMS360 [31]	Zhu *et al.* [33]	Salicon [28]	MLNet [34]	SalGAN [29]	Ours
CC ↑	0.414	0.398	0.427	0.444	0.465	0.499	0.647	0.453	0.629	**0.679**
NSS ↑	0.916	0.880	0.902	0.998	1.096	1.114	1.464	1.031	1.417	**1.621**
AUC ↑	0.743	0.734	0.743	0.760	0.759	0.768	0.830	0.764	0.817	**0.833**
SIM ↑	0.452	0.444	0.452	0.462	0.468	0.464	0.571	0.476	0.576	**0.607**
KLD ↓	4.428	4.620	4.575	4.288	4.040	4.406	2.303	3.490	2.277	**1.907**

Table 2. Quantitative comparison results of different models for saliency prediction on TD group (Learnable models are fine-tuned on our dataset).

Metric\Model	Itti	GBVS [30]	GBVS360 [31]	BMS [32]	BMS360 [31]	Zhu *et al.* [33]	Salicon [28]	MLNet [34]	SalGAN [29]	Ours
CC ↑	0.443	0.425	0.459	0.478	0.493	0.528	0.678	0.535	0.669	**0.724**
NSS ↑	1.007	0.954	1.039	1.104	1.202	1.204	1.286	1.264	1.543	**1.797**
AUC ↑	0.764	0.752	0.764	0.778	0.781	0.790	0.841	0.787	0.834	**0.845**
SIM ↑	0.475	0.465	0.480	0.483	0.490	0.491	0.611	0.522	0.608	**0.641**
KLD ↓	3.915	4.073	3.981	3.883	3.835	3.912	1.828	2.786	1.759	**1.527**

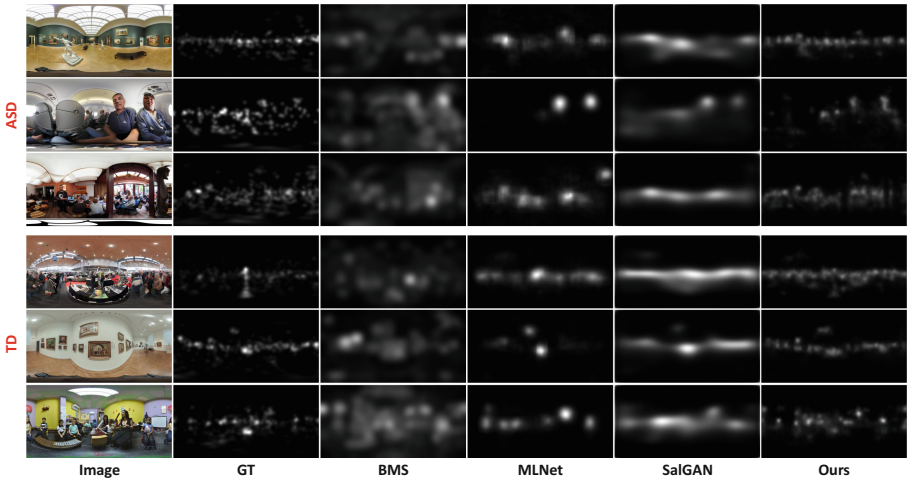

Fig. 6. Qualitative comparisons between different methods.

4 Experimental Results

4.1 Performance Evaluation on Our Dataset

We evaluate the performance of VQSAL on modeling the visual attention of different groups on our dataset. Table 1 shows the quantitative comparisons of different models for modeling the visual attention of children with ASD. It can be observed that our method acquires the best performance compared to other 9 SOTA models. As demonstrated in Table 2, towards modeling the visual attention of TD children, VQSAL also achieves the SOTA results compared with other models.

Table 3. Quantitative comparison results of different models for saliency prediction on other datasets.

Categories	Metric\Model	GBVS360 [31]	BMS [32]	BMS360 [31]	Zhu *et al.* [33]	Salicon [28]	MLNet [34]	SalGAIL [35]	Ours
Overall	CC ↑	0.590	0.557	0.714	0.727	0.511	0.589	0.742	**0.816**
	NSS ↑	0.995	0.975	1.378	1.295	0.856	1.064	1.556	**1.591**
	AUC ↑	0.766	0.758	0.841	0.821	0.757	0.784	0.853	**0.870**
	KLD ↓	0.566	0.584	0.584	0.420	0.637	0.844	0.345	**0.251**

Moreover, we further compare the qualitative results between different models for modeling the visual attention of ASD children and TD children, respectively, as shown in Fig. 6. We can observe that the predicted saliency results of our method are more consistent with the ground-truth (GT). More importantly, compared with other SOTA methods, our method can better describe the local visual attention in detail while other models can only generate a rough visual attention map over the space.

4.2 Generalization Ability on Other Datasets

VQSAL can not only be used to model the visual attention of human groups with different cognitive conditions, but also be extended to the general omnidirectional saliency prediction task. Here we demonstrate the superiority of VQSAL on another omnidirectional saliency prediction database [35]. As shown in Table 3, the overall performances of VQSAL across different metrics are better than other 7 SOTA models.

5 Discussion and Conclusion

In this paper, we present an important problem *i.e.*, where are the children with autism looking in reality? Although there were many previous studies discussing the visual attention of individuals with autism, the vast majority of prior studies not only used stimuli with restricted FOV, but also conducted experiments with relatively small and fixed FOV. To the best of our knowledge, there is no previous study conducting large-scale controllable experiments towards modeling the visual attention of children with autism in reality. Considering two factors, *i.e.*, omnidirectional free viewing and controllable experimental condition, are required to be balanced, we conduct the first large-scale visual attention study under VR environment, towards better understanding and modeling of the gaze pattern of children with autism in reality/VR.

Besides the contribution of the large-scale eye-tracking study and the database, we also apply a saliency prediction method for better modeling the human visual attention. We surprisingly find the consistence between the model and various human groups with different cognitive conditions. Quantitative and qualitative comparisons with SOTA models demonstrate the superiority of this method. Moreover, this method can also be generalized to other omnidirectional saliency prediction datasets and tasks and achieve SOTA performances.

There are many interesting phenomena can be observed from the constructed database. In this paper, though we have analyzed some omnidirectional differences between the visual attention of children with autism and healthy controls, more explorations are needed to further study the common characteristics underlying this complex neurodevelopmental condition in the future. First of all, more statistical analysis are needed to compare the differences between the gaze patterns of two groups, including the quantitative analysis of the fixations and various visual features (*e.g.*, from pixel level features to semantic level features), the relationship between head movement and eye movement, the characteristics of saccades and scanpaths *etc.* Moreover, this study can be seen as a large-scale preliminary research for future works to design specific stimuli towards diagnosis or rehabilitation application development.

References

1. Robertson, C.E., Baron-Cohen, S.: Sensory perception in autism. Nat. Rev. Neurosci. **18**(11), 671 (2017)
2. American Psychiatric Association, et al.: Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub (2013)
3. Simmons, D.R., Robertson, A.E., McKay, L.S., Toal, E., McAleer, P., Pollick, F.E.: Vision in autism spectrum disorders. Vis. Res. **49**(22), 2705–2739 (2009)
4. Wang, S., Xu, J., Jiang, M., Zhao, Q., Hurlemann, R., Adolphs, R.: Autism spectrum disorder, but not amygdala lesions, impairs social attention in visual search. Neuropsychologia **63**, 259–274 (2014)
5. Shi, F., et al.: Drawing reveals hallmarks of children with autism. Displays **67**, 102000 (2021)
6. Osterling, J., Dawson, G.: Early recognition of children with autism: a study of first birthday home videotapes. J. Autism Dev. Disord. **24**(3), 247–257 (1994)
7. Dawson, G., Webb, S.J., McPartland, J.: Understanding the nature of face processing impairment in autism: insights from behavioral and electrophysiological studies. Dev. Neuropsychol. **27**(3), 403–424 (2005)
8. Sasson, N.J., Elison, J.T., Turner-Brown, L.M., Dichter, G.S., Bodfish, J.W.: Brief report: circumscribed attention in young children with autism. J. Autism Dev. Disord. **41**(2), 242–247 (2011)
9. Corden, B., Chilvers, R., Skuse, D.: Avoidance of emotionally arousing stimuli predicts social-perceptual impairment in Asperger's syndrome. Neuropsychologia **46**(1), 137–147 (2008)
10. Klin, A., Jones, W.: Altered face scanning and impaired recognition of biological motion in a 15-month-old infant with autism. Dev. Sci. **11**(1), 40–46 (2008)
11. Jiang, M., Zhao, Q.: Learning visual attention to identify people with autism spectrum disorder. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3267–3276 (2017)
12. Duan, H., et al.: A dataset of eye movements for the children with autism spectrum disorder. In: Proceedings of the ACM Multimedia Systems Conference (MMSys) (2019)
13. Duan, H., et al.: Learning to predict where the children with ASD look. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 704–708 (2018)

14. Duan, H., Min, X., Fang, Y., Fan, L., Yang, X., Zhai, G.: Visual attention analysis and prediction on human faces for children with autism spectrum disorder. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **15**(3s), 1–23 (2019)
15. Fang, Y., Duan, H., Shi, F., Min, X., Zhai, G.: Identifying children with autism spectrum disorder based on gaze-following. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 423–427. IEEE (2020)
16. Bellani, M., Fornasari, L., Chittaro, L., Brambilla, P.: Virtual reality in autism: state of the art. Epidemiol. Psychiatr. Sci. **20**(3), 235–238 (2011)
17. Rai, Y., Gutiérrez, J., Le Callet, P.: A dataset of head and eye movements for 360 degree images. In: Proceedings of the ACM on Multimedia Systems Conference, pp. 205–210 (2017)
18. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2695–2702. IEEE (2012)
19. Duan, H., Zhai, G., Yang, X., Li, D., Zhu, W.: IVQAD 2017: an immersive video quality assessment database. In: Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 1–5 (2017)
20. Zhu, Y., Zhai, G., Yang, Y., Duan, H., Min, X., Yang, X.: Viewing behavior supported visual saliency predictor for 360 degree videos. IEEE Trans. Circuits Syst. Video Technol. (TCSVT) (2021)
21. Duan, H., Shen, W., Min, X., Tu, D., Li, J., Zhai, G.: Saliency in augmented reality. In: Proceedings of the ACM International Conference on Multimedia (ACM MM) (2022)
22. Duan, H., Min, X., Zhu, Y., Zhai, G., Yang, X., Callet, P.L.: Confusing image quality assessment: towards better augmented reality experience. arXiv preprint arXiv:2204.04900 (2022)
23. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2202–2210 (2022)
24. Sitzmann, V., et al.: Saliency in VR: how do people explore virtual environments? IEEE Trans. Vis. Comput. Graph. **24**(4), 1633–1642 (2018)
25. Haskins, A.J., Mentch, J., Botch, T.L., Robertson, C.E.: Active vision in immersive, 360 real-world environments. Sci. Rep. **10**(1), 1–11 (2020)
26. Wang, S., et al.: Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. Neuron **88**(3), 604–616 (2015)
27. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. arXiv preprint arXiv:2012.09841 (2020)
28. Huang, X., Shen, C., Boix, X., Zhao, Q.: SALICON: reducing the semantic gap in saliency prediction by adapting deep neural networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 262–270 (2015)
29. Pan, J., et al.: SalGAN: visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081 (2017)
30. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency (2007)
31. Lebreton, P., Raake, A.: GBVS360, BMS360, ProSal: extending existing saliency prediction models from 2D to omnidirectional images. Signal Process.: Image Commun. **69**, 69–78 (2018)
32. Zhang, J., Sclaroff, S.: Exploiting surroundedness for saliency detection: a Boolean map approach. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **38**(5), 889–902 (2015)
33. Zhu, Y., Zhai, G., Min, X.: The prediction of head and eye movement for 360 degree images. Signal Process.: Image Commun. **69**, 15–25 (2018)

34. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), pp. 3488–3493 (2016)
35. Xu, M., Yang, L., Tao, X., Duan, Y., Wang, Z.: Saliency prediction on omnidirectional image with generative adversarial imitation learning. IEEE Trans. Image Process. (TIP) **30**, 2087–2102 (2021)