Pointing Gesture Recognition via Self-supervised Regularization for ASD Screening

Cheol-Hwan Yoo, Jang-Hee Yoo, Ho-Won Kim, and ByungOk Han
ETRI
Daejeon, Republic of Korea
{ch.yoo, jhy, hw_kim, byungok.han} @etri.re.kr

Abstract

The ability to point to objects for sharing social purpose or attention is known as one of the key indicators in distinguishing children with typical development (TD) from those with autism spectrum disorder (ASD). However, there is a lack of datasets specifically tailored for children's pointing gestures. This lack of training data from the target domain becomes a major factor in the performance degradation of conventional supervised CNNs due to domain shift. Toward an effective and practical solution, we propose an end-to-end learning scheme for domain generalized pointing gesture recognition adopting self-supervised regularization (SSR). To prove the effectiveness of our method in real-world situations, we designed a Social Interaction-Inducing Content (SIIC)-based ASD diagnostic system and collected an ASD-Pointing dataset consisting of 40 TD and ASD children. Through extensive experiments on our collected datasets, we achieved an ASD screening accuracy of 72.5%*, showing that pointing ability can play a vital role as an indicator in distinguishing between ASD and TD.*

1. Introduction

Social intelligence, a fundamental aspect of human interaction, is known to be significantly impaired in individuals diagnosed with autism spectrum disorder (ASD) [4, 17]. Early diagnosis of ASD is critical in that the brain of infants and toddlers is highly plastic providing an opportunity to change to a normal form, as well as preventing secondary neurological damage and the accumulation of serious behavior problems. However, the current diagnosis system mainly relies on labor-intensive manual examinations performed by medical experts, which causes a problem of missing early diagnosis, a crucial factor in prognosis. To alleviate these problems, recent studies on the computer-aided diagnosis of ASD based on several signs such as gripping motions, repetitive behaviors, eye movement, aberrant gait, and facial traits have gained significant research interest [22–24, 32, 33, 36, 43]. In addition to these indicators, pointing skills that first occur between the ages of 8 and 10 months and account for the majority of gestures [5] can also be a crucial indicator in the early diagnosis of ASD. In general, pointing can be categorized into two types: *Protoimperative* and *protodeclarative*. Among these types of pointing, *protodeclarative* pointing has the purpose of sharing social attention or interest with others, and thus its deficits can be an important diagnostic criterion in ASD screening [3, 8, 13, 25, 29].

Based on this medical research, in this paper, we devised a deep learning-based automatic diagnostic system to recognize the pointing ability of children during the Social Interaction-Inducing Content (SIIC)-based test for ASD screening. Since few datasets can directly handle pointing gesture recognition of children, we propose to address this problem by leveraging recent studies on self-supervised learning (SSL) [8–11, 19, 20, 39, 40] and domain generalization (DG) [7, 26]. To this end, we propose an end-to-end deep learning framework for domain-generalized pointing gesture recognition adopting self-supervised regularization (SSR) where domain-agnostic models are trained only on source domains and generalize to unseen test domains. Finally, to validate the effectiveness of our proposed method, we collected and tested real-world *ASD-Pointing* datasets from 40 subjects composed of TD and ASD children which have not been directly used for training. Our contributions can be summarized as follows:

1) To the best of our knowledge, we are the first to cast the problem of aiding an early diagnosis of ASD based on nonverbal behavior, especially children's pointing abilities using deep learning frameworks.

2) We propose a domain-generalized pointing gesture recognition scheme with SSR so that the deep models learn domain-invariant features, which can be applied to various fields such as medical AI with limited access to large datasets and fine-grained annotations.

3) We designed SIIC-based tests and collected real-world *ASD-Pointing* datasets composed of TD and ASD children with video-level annotations from 40 individual subjects. The dataset not only can be used to prove experimentally that pointing ability is closely related to the early diagnosis of ASD but also be expected to be valuable to the community involved.

2. Proposed Method

2.1. Description of the proposed method

Fig. 1 shows the overall architecture of the proposed framework. The proposed framework in the training stage consists mainly of four major parts: a hand detector that crops the person's hand region to discard unnecessary information such as background and body features; multiple encoders with an identical network architecture that extract latent feature embeddings; a self-supervised regularization block (SRB) that makes the architecture asymmetric to prevent collapse; a logit layer that outputs the raw probability values for classification results. In the test stage, only the single branch of the network is used to recognize pointing gestures with an additional ensemble block to add the robustness of network predictions.

Following recent studies exploring the importance of body parts in action and attribute classification [18], our framework firstly detects a person's hand using a hand detector $\mathcal{D}(\cdot)$ from each frame x of the input videos which drives the network to only focus on the hand region by discarding unnecessary background and body features. Given a detected hand region, $\mathcal{D}(x)$ of the input frame, k-th randomly augmented views x_k are generated as follows:

$$x_k = \mathcal{T}_k(\mathcal{D}(x)), \tag{1}$$

where $\mathcal{T}_k(\cdot)$, $k = 1, \cdots, N$ represents the transformation function.

To learn domain-invariant features, these randomly augmented images are fed into the encoders followed by SRB that makes the distance between randomly augmented images become close to each other as follows:

$$f_k = \mathcal{E}_k(x_k; \Theta_e^k), \tag{2}$$

$$L_{reg} = \text{SRB}(f_1, \cdots, f_k), \tag{3}$$

where $\mathcal{E}_k(\cdot)$, $k = 1, \cdots, N$, represents an encoder of the network with trainable Θ_e^k parameters, f_k, $k = 1, \cdots, N$, are the latent feature embeddings, and L_{reg} is the self-supervised regularization (SSR) loss induced from SRB, respectively. Note that the number of applied transformations, N, can be extended to arbitrary sizes, but following most self-supervised learning methods [9, 10, 19, 20, 40], it is set to two in our papers.

We also adopt the binary cross-entropy loss function to the first branch of the network for training the task of binary classification, i.e. pointing or no-pointing, as follows:

$$y_1 = \mathcal{P}(\mathcal{E}_1(x_1; \Theta_e^1); \Theta_{\mathcal{P}}), \tag{4}$$

$$L_c = -\sum_{i}^{2} t_i log(s(y_1))_i, \tag{5}$$

where $\mathcal{P}$ is the logit layer with trainable parameters Θ_P, s is the softmax activation function, and t_i denotes the i-th component of one-hot ground-truth vector t, respectively.

Thus, we define the total loss functions for jointly training our network which consists of classification loss L_c and SSR loss L_{reg} as follows:

$$L = L_c + \lambda L_{reg}, \tag{6}$$

where λ is a weight parameter to control the two losses. For L_{reg}, it can vary depending on which SSL method is selected. For example, when selecting SimSiam [10] as the SRB, the loss function L_{reg} with the negative cosine similarity D_{cos} is defined as:

$$D_{cos}(f_1, h_2) = -\frac{f_1}{\|f_1\|_2} \cdot \frac{h_2}{\|h_2\|_2}, \tag{7}$$

$$L_{reg} = \frac{1}{2}(D_{cos}(sg(f_1), h_2) + D_{cos}(h_1, sg(f_2))), \tag{8}$$

where $\|\cdot\|_2$ denotes l_2-norm, f and h represent a latent feature and its transformed feature from predictor, respectively. $sg(\cdot)$ denotes the stop-gradient for preventing collapse.

Unlike common approaches where the network is pre-trained with a self-supervised learning (SSL) scheme and fine-tuned to learn the linear classifier [9, 19, 40], we utilize the SSR as an auxiliary constraint to drive the network to learn not only class-specific but also domain-invariant feature representations. Furthermore, the proposed learning framework can be compatible with any existing SSL methods, and the whole network can be easily trained in an end-to-end manner.

2.2. Ensemble block

In the test stage, to reduce the risk of per-frame prediction which is vulnerable to noise, and incorporate the temporal information from input sequences, we adopt a simple voting scheme called ensemble block as in [2]. The ensemble block takes current and previous frame-level predictions using a temporal sliding window. To predict the final video-level result for pointing gestures, each frame-level prediction is aggregated as:

$$y = \mathcal{A}(y_1^t, y_1^{t-1}, \cdots, y_1^{t-T}), \tag{9}$$

where $\mathcal{A}(\cdot)$ is a temporal ensemble block, t is the current time step, and T is the number of previous frames within a

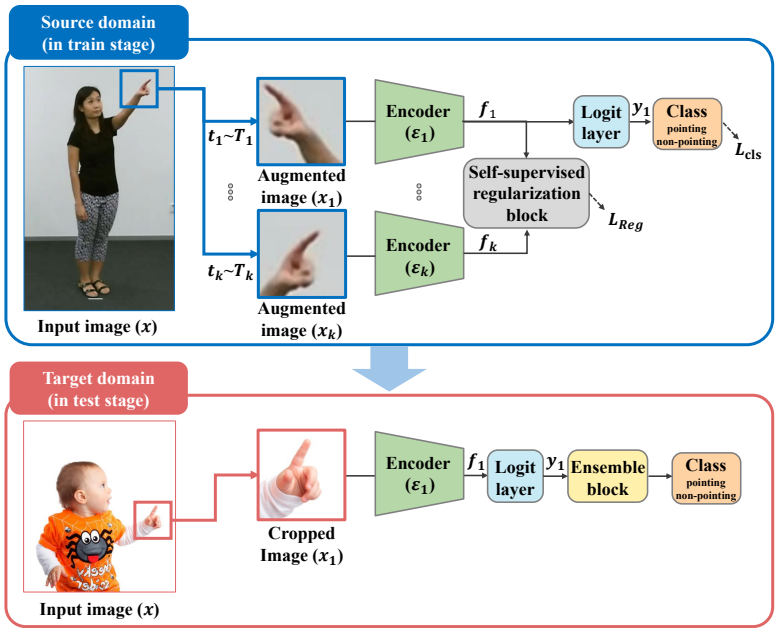

Figure 1. The overall framework of the proposed method for pointing gesture recognition.

Datasets	NTU RGB+D
No. of Images(train)	48.7k
No. of Images(val)	12.2k

Table 1. Configurations of re-organized and re-purposed datasets for training the proposed network.

Groups	Gender		Age (years)				
	Male	Female	1-2	2-3	3-4	4-5	5-6
ASD($N = 26$)	20	6	1	3	12	10	0
TD($N = 14$)	4	10	4	1	3	2	4

Table 2. Subjects, gender, and age distributions for *ASD-Pointing* dataset.

sliding window which is set to two in our paper. In other words, when all the recent three consecutive frames are agreed upon, it is determined that a final positive pointing reaction has occurred.

2.3. Train dataset construction

As far as we know, most of the publicly available datasets for human action or hand gesture recognition, including pointing classes, cannot be directly used to train pointing gesture recognition networks with binary annotation (pointing or no-pointing). Therefore, we re-purposed and re-organized existing action recognition datasets, NTU RGB+D [34] for training our proposed network. The NTU RGB+D dataset contains 60 action classes and 56,880 video samples. Among the NTU RGB+D dataset, the hand regions obtained from *point to something* class were defined as positive samples, and the hand regions randomly obtained as many as the positive sample in the remaining action classes were defined as negative samples (i.e. One-vs-Rest strategy). To detect the hand region, we use a hand detector of MMPose (cascade rcnn x101 64x4d fpn 1class) [14], and the detected hand region is resized to the size of 256×256. The configurations of the dataset are described in Table 1 and we will make the dataset publicly available.

2.4. ASD-Pointing dataset construction

To validate the effectiveness of our method in ASD screening, we designed and developed a Social Interaction-Inducing Content (SIIC)-based diagnostic system. The system aims to prompt well-known social behaviors associated with early signs of ASD children, such as joint attention, eye contact, social smile, pointing gestures, and response to name-calling. The SIIC-based system uses four Microsoft Azure Kinect cameras to record children's social interactions while they watch the SIIC being played on three monitors as shown in Fig. 2. Fig. 3 shows examples of the SIIC for inducing pointing gestures of children. We constructed an *ASD-pointing* dataset using recorded video clips during the pointing gesture induction intervals in the SIIC. The length of the content for pointing gestures induction is 5 seconds, and it's repeated three times for each subject with different instructions (e.g. *Look for a tiger, apple, and airplane*). Fig. 4 shows examples of captured images through our SIIC-based system from four different viewpoints. In addition to ASD diagnosis, the presence or absence of a positive response to a pointing gesture in the SIIC-based examination is annotated by medical experts after all examinations have been done. Therefore, the final constructed

Figure 2. SIIC-based diagnostic system that triggers social behaviors related to early signs of ASD children in three living lab spaces.

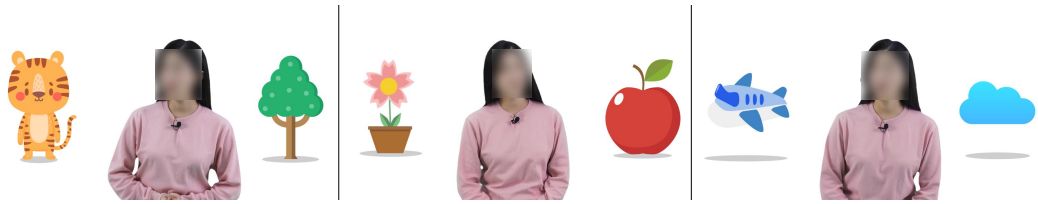

Figure 3. Example images of SIIC-based tests for inducing pointing gestures of children.

dataset contains 480 test video clips from 40 subjects, including 26 ASD and 14 TD children, from three living lab spaces. The details of the *ASD-Pointing* dataset are described in Table 2. All the above studies were approved by Institutional Review Board at Seoul National University Hospital and Pusan National University and were signed by all participants on a consent form with detailed descriptions of the research.

3. Experiments

3.1. Implementation details

We adopt ResNet-50 [21] and Vision Transformer (ViT-B/32) [16] as the base network for the encoder with the ImageNet-1K weight initialization. For training our networks, we use the SGD optimizer with an initial learning rate of $1e-4$, and a weight decay of $1e-4$ with a batch size of 8. The learning rate is multiplied by a factor of 0.1 after every 20 epoch. Our model is trained for 50 epochs with a single NVIDIA RTX A6000. Then, the best model on the validation set during the training is chosen for final testing. For a weight factor λ, we set it to 0.5 which shows the best results. During training of our proposed networks, each input image is transformed twice to generate augmented images with random cropping of output size 224, and horizontal flipping. For the baseline without SSR, the same augmentations on the input images are applied. In the test stage on the real-world *ASD-Pointing* dataset, to identify children among the people present in the scene and robustly detect the hand region in practical situations, we utilize OpenPose [6] with depth information from the Kinect sensor instead of using MMPose as in the training stage. Specifically, we lifted the 2D body coordinates obtained by OpenPose to 3D coordinates using the depth information and camera parameters given in the Kinect sensor. The 3D bone length between shoulders is calculated, and the person with the shortest bone length in the image is selected as a baby which is a target person in our experiment. To crop the hand region, following the method in [35], we approximated the hand position using the elbow and wrist position, assuming that the hand is about half the length of the forearm in the same direction. A fixed-size cube of size $150mm$ around the hand location is extracted, projected into the 2D image space, and then resized to the size of 224×224 image. For each frame, hand gesture recognition is performed on both detected hands, and the final predictions are made through OR operation.

Crop224	H-Flip	ColorJitter	Grayscale	G-Blur	Acc(%)
✓	✓				**86.5**
✓	✓	✓			83.1
✓	✓		✓		86.3
✓	✓			✓	84.2
✓	✓	✓	✓	✓	76.5

Table 3. Ablation study on the effect of adding data augmentation technique on the *ASD-Pointing* dataset.

3.2. Ablation studies

In general, SSL methods are known to be dependent on the choice of data augmentation. To select the best combination of data augmentation in our proposed method, following SimSiam [10], we experimented with the data augmentation techniques: random cropping of size 224, random horizontal flipping, random color jittering, random grayscaling, random Gaussian blurring. For our ablation studies, we select SimSiam as the SRB with ResNet-50 backbone. As shown in Table 3, the best results are achieved when the random crop of size 224, and horizontal flip (i.e. weak augmentations) were used. In contrast to results reported in papers [19, 40], the classification accuracy

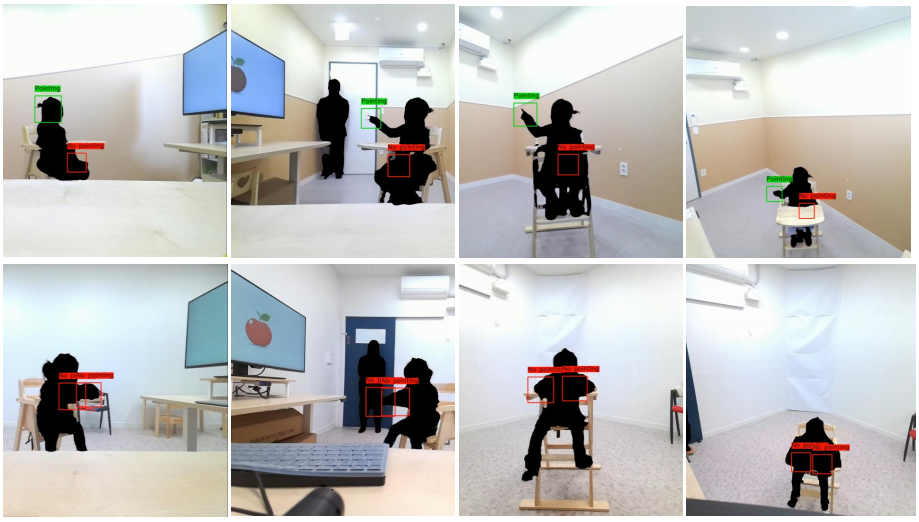

Figure 4. Examples of result images from different camera views where pointing is performed (top row) or not (bottom row) during the SIIC-based testing. Due to privacy issues, the human area was blanked through segmentation.

Model	NTU RGB+D → *ASD-Pointing*			
	Accuracy	Recall	Precision	F1-score
ResNet-50 (baseline)	84.8	66.2	53.5	59.2
Proposed$_{\text{BYOL}}$	<u>86.3</u>	63.7	58.0	<u>60.7</u>
Proposed$_{\text{SimSiam}}$	**86.5**	76.2	57.0	**65.2**
ViT-B/32 (baseline)	61.9	98.7	30.3	46.3
Proposed$_{\text{BYOL}}$	<u>75.0</u>	97.5	39.8	<u>56.5</u>
Proposed$_{\text{SimSiam}}$	**75.2**	98.7	40.1	**57.0**

Table 4. Performance evaluation of pointing gesture recognition on real-world ASD-pointing datasets. Bold and underline indicate the best and second-best results, respectively.

was rather dropped when random color jittering, random grayscaling, and random Gaussian blurring (i.e. strong augmentations) were added. We speculate that these results can be attributed to the distortion of the inherent appearance of training images caused by strong augmentation techniques which could potentially harm the training process. Following the results, unless otherwise noted, we used random cropping and horizontal flipping in the rest of the papers.

3.3. Performance evaluation of pointing gesture recognition on real-world ASD-Pointing datasets

The results of pointing gesture recognition on the *ASD-Pointing* datasets in terms of average accuracy, precision, recall, and F1-score are listed in Table 4. Note that we only used re-organized NTU RGB+D dataset described in Section 2.3 for training, and tested on unseen *ASD-Pointing* dataset to validate the generalization ability of the proposed method. Any existing SSL methods can be adopted in our frameworks, but for applicability and scalability, we adopt SimSiam [10] and BYOL [19] in SRB which do not require negative samples. As shown in the results, adopting our proposed learning scheme on the baseline network boosts the overall classification performance. As for the ResNet-50 backbone, the Proposed$_{\text{SimSiam}}$ and Proposed$_{\text{BYOL}}$ improve the accuracy of Vanilla ResNet-50 by 1.7%p and 1.5%p, respectively. As for the transformer-based architecture, the overall accuracy of the networks using the ViT backbone is lower than that of the ResNet-50. We speculate that this is due to the difficulty of learning for transformer networks that rely on a large amount of training data in the lack of inductive bias. Nevertheless, the Proposed$_{\text{SimSiam}}$ and Proposed$_{\text{BYOL}}$ reduce the domain gap and dramatically improve the accuracy of the Vanilla ViT network by 13.3%p and 13.1%p, respectively. The experimental results demonstrate that adopting our proposed learning scheme with SSR leads to a domain-robust feature representation that consistently improves the generalization ability of the deep models without depending on a particular choice of backbone networks or SSL methods.

3.4. Screening of ASD children based on the pointing ability

We present the results of screening ASD children based on their pointing ability during the SIIC-based test in Table 5. For the test, we use the Proposed$_{\text{SimSiam}}$ with ResNet-50 backbone which shows the best performance on pointing gesture recognition. Each column marked 1^{st}, 2^{nd}, and 3^{rd} denotes the average of the pointing probability (i.e. softmax prediction values) from four camera views for each of the three pointing activities. As can be seen in the results, ASD children had less positive responses to pointing gestures than TD children, and only one in 26 ASD children recorded a higher pointing probability value than 50. Table 6 shows the screening results of ASD/TD based on the

Subjects		1st	2nd	3rd	Avg.(%)	Subjects		1st	2nd	3rd	Avg.(%)
ASD	1	83.3	84.3	40.1	**69.2**	ASD	22	5.8	35.4	66.5	35.9
	2	41.8	35.4	18.9	32.0		23	1.4	0.6	21.6	7.8
	3	2.3	16.8	21.3	13.5		24	4.5	21.8	5.1	10.5
	4	2.3	1.4	3.2	2.3		25	1.3	1.5	1.3	1.3
	5	43.0	19.9	3.0	22.0		26	2.8	4.6	29.1	12.1
	6	1.5	4.7	66.6	24.3	Avg. probability		12.2	17.2	18.7	16.1
	7	17.9	22.5	18.4	19.6	TD	1	62.2	84.3	60.0	**68.9**
	8	2.8	4.2	90.5	32.5		2	21.0	5.2	3.5	9.9
	9	2.8	3.5	1.4	2.6		3	66.7	93.6	75.9	**78.7**
	10	2.1	2.8	3.0	2.6		4	74.4	78.9	68.3	**73.8**
	11	23.6	45.3	21.8	30.2		5	77.7	16.4	19.6	37.9
	12	4.3	5.4	2.7	4.1		6	73.4	89.2	67.9	**76.8**
	13	1.5	0.7	14.7	5.6		7	2.1	2.8	2.6	2.5
	14	1.5	17.0	3.1	7.2		8	2.5	2.9	3.0	2.8
	15	2.6	3.0	2.7	2.8		9	1.0	19.3	6.1	8.8
	16	2.2	1.5	1.6	1.7		10	0.9	1.3	1.1	1.1
	17	3.8	3.5	5.7	4.3		11	3.7	1.4	1.3	2.1
	18	3.6	47.1	18.6	23.1		12	21.7	55.8	3.2	26.9
	19	3.5	4.1	2.1	3.2		13	6.7	6.5	6.0	6.4
	20	53.7	58.5	21.8	44.6		14	1.1	0.9	0.9	0.9
	21	1.7	3.0	2.7	2.5	Avg. probability		29.6	32.7	22.8	28.4

Table 5. Probability of positive pointing response during the content-based tests.

N=40	ASD	TD	Predicted
ASD	25	1	Recall: 96.2%
TD	10	4	-
Actual	Precision: 71.4%	-	Accuracy: 72.5%

Table 6. Confusion matrix for ASD screening, with a probability threshold of 50.

probability threshold of 50. Here we obtained screening accuracy, recall, precision, and F1-score of 72.5%, 96.2%, 71.4%, and, 82.0%, respectively. The ASD screening accuracy of 72.5% with only a short test time of about 15 seconds in the SIIC-based test proves a high potential of pointing ability to be used as a key indicator in discriminating between ASD and TD children. Furthermore, given the characteristics of the medical field where classifying positive samples as negative has more risk than the opposite case, the relatively high recall rate of 96.2% can be considered to have the advantage of being able to be used as a primary screening tool.

4. Conclusion

In this paper, we explored a method for predicting the diagnosis of children with ASD and TD based on the ability to perform pointing gestures through the Social Interaction-Inducing Content-based test. Toward a practical approach in the field of medical AI where access to large datasets is limited, we designed a simple but effective training scheme with self-supervised regularization. We also collected a real-world *ASD-Pointing* dataset from 40 subjects composed of 14 TD and 26 ASD children. Our experiments on *ASD-Pointing* dataset show that adopting the proposed learning scheme improves the generalization ability of the base networks, alleviating the challenge of the unavailability of datasets. Furthermore, a screening accuracy of 72.5% based on the pointing ability during a short testing time of 15 seconds indicates a high potential for pointing gestures to be used as a key indicator in discriminating between ASD and TD children.

5. Limitation

1) Screening ASD with only one behavioral indicator has a limitation in reliability, even if it shows a certain degree of accuracy. Even TD children may not perform behaviors that are a target indicator due to unfamiliar test environment, and on the contrary, ASD children may have the ability to perform a specific behavior. Combining key features that distinguish children with ASD from TD children, such as joint attention, social smiling, eye contact, response to name calling, and repetitive behavior can not only increase the accuracy of ASD screening but also make a high reliable diagnosis.
2) As shown in the result in Section 3.4, most of the ASD children (14 out of 15 children) did not perform pointing, but even some cases of TD children also did not point. We speculate that the unfamiliar testing environment may have hindered TD children from making positive responses to pointing gestures, which is expected to be addressed in our future work through improvements to SIIC-based tests, such as adding a warming-up section.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2019-0-00330, Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders based on Cognition of the Psychological Behavior and Response)

References

[1] Dosovitskiy Alexey, Philipp Fischer, Jost Tobias, Martin Riedmiller Springenberg, and Thomas Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 99, 2015.

[2] Oscar L Barbed, Pablo Azagra, Lucas Teixeira, Margarita Chli, Javier Civera, and Ana C Murillo. Fine-grained pointing recognition for natural drone guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1040–1041, 2020.

[3] Simon Baron-Cohen. Perceptual role taking and protodeclarative pointing in autism. *British journal of developmental psychology*, 7(2):113–127, 1989.

[4] Simon Baron-Cohen, Howard A. Ring, Sally Wheelwright, Edward T. Bullmore, Mick J. Brammer, Andrew Simmons, and Steve C. R. Williams. Social intelligence in the normal and autistic brain: an fmri study. *European Journal of Neuroscience*, 11(6):1891–1898, 1999.

[5] Elizabeth Bates and Frederic Dick. Language, gesture, and the developing brain. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 40(3):293–310, 2002.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[7] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[12] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021.

[13] Caitlin Clements and Katarzyna Chawarska. Beyond pointing: Development of the "showing" gesture in children with autism spectrum disorder. *Yale Review of Undergraduate Research in Psychology*, 2:1–11, 2010.

[14] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Fifth Edition et al. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643, 2013.

[18] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Actions and attributes from wholes and parts. In *Proceedings of the IEEE international conference on computer vision*, pages 2470–2478, 2015.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Shashank Jaiswal, Michel F Valstar, Alinda Gillott, and David Daley. Automatic detection of adhd and asd from expressive behaviour in rgbd data. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 762–769. IEEE, 2017.

[23] M Jazouli, S Elhoufi, A Majda, A Zarghili, and R Aalouane. Stereotypical motor movement recognition using microsoft kinect with artificial neural network. *International Journal of Computer and Information Engineering*, 10(7):1270–1274, 2016.

[24] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *Proceedings of the ieee international conference on computer vision*, pages 3267–3276, 2017.

[25] Chris Plauché Johnson, Scott M Myers, et al. Identification and evaluation of children with autism spectrum disorders. *Pediatrics*, 120(5):1183–1215, 2007.

[26] Daehee Kim, Youngjun Yoo, Seunghyun Park, Jinkyu Kim, and Jaekoo Lee. Selfreg: Self-supervised contrastive regularization for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9619–9628, 2021.

[27] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[28] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.

[29] Filippo Muratori, L Camaioni, P Perucchini, and A Milone. A longitudinal examination of the communicative gestures deficit in young children with autism. 1997.

[30] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[31] Mehdi Noroozi, Hamed Pirsiavash, and Paolo Favaro. Representation learning by learning to count. In *Proceedings of the IEEE international conference on computer vision*, pages 5898–5906, 2017.

[32] Shahid Omar, Sejuti Rahman, Faiza Ahmed Syeda, Musabbir Ahmed Arrafi, and MAR Ahad. Data-driven automated detection of autism spectrum disorder using activity analysis: A review. *Preprints*, 2020.

[33] Omar Rihawi, Djamal Merad, and Jean-Luc Damoiseaux. 3d-ad: 3d-autism dataset for repetitive behaviours with kinect sensor. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[34] 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*, pages 1010–1019, 2016.

[35] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

[36] Uzma Haque Syeda, Ziaul Zafar, Zishan Zahidul Islam, Syed Mahir Tazwar, Miftahul Jannat Rasna, Koichi Kise, and Md Atiqur Rahman Ahad. Visual face scanning and emotion perception analysis between autistic and typically developing children. In *Proceedings of the 2017 acm international joint conference on pervasive and ubiquitous computing and proceedings of the 2017 acm international symposium on wearable computers*, pages 844–853, 2017.

[37] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *Advances in neural information processing systems*, 31, 2018.

[38] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021.

[39] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers. *arXiv preprint arXiv:2105.04553*, 2021.

[40] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[41] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020.

[42] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[43] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, Jessica Podda, Francesca Battaglia, Edvige Veneselli, Cristina Becchio, and Vittorio Murino. Video gesture analysis for autism spectrum disorder detection. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3421–3426. IEEE, 2018.