Technique Report │ Received 4 November 2024; Accepted 17 February 2025; Published 25 February 2025
https://doi.org/10.55092/bi20250002

Micro-expression detection in ASD movies: a YOLOv8-SMART approach

Yutong Gu[1], Hanni Li[1], Jiarong Liu[1], Chenxi Liu[1], Yuxuan Li[1], Chen Li[1],* and Ning Xu[2],*

[1] School of Biomedical Engineering, Northeastern University, Shenyang, China
[2] School of Art and Design, Liaoning Petrochemical University, Fushun, China

* Correspondence authors; E-mails: lichen@bmie.neu.edu.cn (C.L.); xuning201096@hotmail.com (N.X.).

Highlights:

- **ASD micro-expression dataset:** A unique dataset of ASD patients' micro-expressions in movies is used to support early diagnosis.
- **YOLOv8-SMART algorithm:** An improved YOLOv8 algorithm enhances micro-expression detection accuracy in ASD patients.
- **Early diagnosis potential:** The method provides a practical tool for doctors to improve ASD recognition and intervention through micro-expression analysis.

Abstract: *Autism Spectrum Disorder (ASD)* is a neurodevelopmental disorder in which individuals often face social difficulties as well as language and communication challenges. Micro-expressions are extremely brief changes in facial expression. Moreover, the micro-expressions exhibited by individuals with ASD frequently represent an accurate reflection of their internal feelings. Therefore, using the Cinemetrics method to extract micro-expressions from ASD patients in movies and targeting them for detection can help doctors make early diagnosis of ASD patients. In this paper, we establish a dataset of micro-expressions of ASD patients in movies, use the improved YOLOv8-SMART algorithm for target detection, and compare it with other target detection algorithms without improvement. The comparison results prove that our algorithm effectively improves the recognition of micro-expressions, which provides reference value for future practical applications in the task of micro-expression recognition in ASD patients.

Keywords: autism spectrum disorder; Cinemetrics; micro-expressions; movies; YOLOv8-SMART; target detection algorithms

1. Introduction

Autism spectrum disorder (ASD) is a neurological and developmental syndrome rather than a single disorder [1]. In recent years, the prevalence of ASD has risen significantly, affecting approximately 1 in 100 children and adolescents [2,3], although data on prevalence in low-income countries remain limited [4]. ASD is characterized by challenges in language, social cognition, and psychiatric functioning [1], often identified in early childhood, yet diagnosis can be delayed. Currently, as noted by Eissa *et al.*, the only drugs approved for clinical use to alleviate ASD symptoms are risperidone and aripiprazole, which primarily address behavioral issues [5]. Consequently, rehabilitation remains the mainstay of treatment. It is essential to understand and capture the emotional experiences of individuals with ASD to support their recovery effectively.

Micro-expressions are very brief, involuntary facial expressions [6] that have the potential to reveal hidden emotions. In the 1960s, this hidden change in facial emotion was discovered by a journalist named Ekman. In the interview, a video of a mentally ill patient played in slow motion showed a super brief expression of sadness-lasting only 1/12th of a second [7]. The micro-amplitude and rapidity of facial movements make it difficult for the human eye and even experienced doctors and specialists to recognize micro-expressions in people with ASD [8]. Consequently, a real-time micro-expression analysis system can enhance the ability of individuals with ASD to comprehend genuine emotional responses, thereby facilitating timely diagnosis and treatment.

Cinemetrics is a tool and methodology for analyzing the quantitative features of movies [9]. This innovative approach fosters a deeper exploration of cinematic language, character relationships, and affective expressions. It automatically collects data from various dimensions of a movie, including shot length, editing speed, scene changes, and audio characteristics [10]. By employing a Cinemetrics-based approach, micro-expressions of individuals with ASD in movies can be extracted and analyzed, thereby providing objective data that supports clinical diagnosis and aids in the early identification of ASD patients.

In this study, we constructed a dataset based on metric cinematography, utilizing character micro-expression labels from ASD movies. We adopted the improved YOLO-V8 model as the foundation for an automatic micro-expression analysis system to enhance the recognition of emotional states in ASD patients. The improved YOLO-V8 model significantly increases the recognition accuracy of complex facial micro-expressions through the in-depth extraction of subtle features. Additionally, the model supports multi-scale detection, allowing it to effectively recognize changes in micro-expressions across varying distances and angles. Figure 1 illustrates the specific workflow of this study.

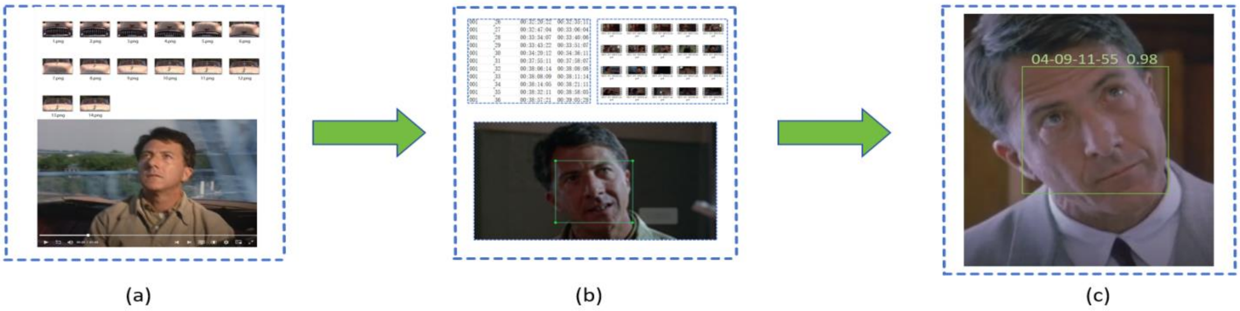


Figure 1. Specific workflow of our study: **(a)** Data collecting: extract relevant clips featuring ASD characters from the corresponding movies; **(b)** Data labeling: assign micro-expression labels to the characters; **(c)** Micro-expression recognition: accurate sentiment analysis is achieved using the trained model [11].

2. Related Work

2.1. Research on ASD patients

ASD is a type of neurobehavioral disorder characterized by genetic heterogeneity, with primary features including difficulties in communication, social interaction, and repetitive behaviors [12]. In 1943, Leo Kanner first introduced the term "Autism" to describe the social and emotional relationship difficulties observed in young children [13]. The etiology of ASD involves the interplay of various genetic variants and environmental factors. For instance, maternal use of antidepressants during pregnancy may be linked to an increased risk of ASD in the fetus [14]. According to the American Psychiatric Association, the core symptoms of ASD include difficulties in social communication, sensory abnormalities, and repetitive movements or interests [15]. Additionally, many individuals with ASD experience co-occurring mental health issues such as anxiety and depression. As a result, treatment plans for ASD patients often need to address these accompanying mental health concerns in an integrated manner to develop more effective intervention strategies.

2.2. Micro-expression recognition

Micro-expressions are unconscious facial movements that usually occur when a person experiences an emotion but tries to hide their true underlying emotion [16]. The phenomenon of micro-expressions was first discovered by Haggard and Isaacs in 1966 [17]. Soon after, Ekman *et al.* continued to study micro-expressions and developed the Facial Action Coding System (FACS). FACS decomposes micro-expressions into different action units (AUs), *i.e.*, based on different components of the muscles [18]. Analysis by standardized method AU coding improves the accuracy of recognizing micro-expressions and helps to decode specific emotional responses.

Due to its practical importance in multimodal fusion, micro-expression recognition has attracted increasing attention. The earliest research was carried out using hand-crafted feature engineering techniques. The study was conducted by Pfister *et al.* on the first public spontaneous ME dataset-SMIC using Local Binary Patterns from three orthogonal planes (LBP-TOP) [19]. After that, Li and his team [20] performed micro-expression recognition by fusing LBP-TOP, orientation gradient histograms, and image gradient orientation histograms, which in turn produced individual feature vectors.

Soon after, the development of deep learning techniques significantly boosted the research and application of micro-expression recognition. Patel and colleagues [21] took an innovative approach in the field of micro-expression recognition by using a convolutional neural network (CNN) to train a dataset as a feature extractor. The extracted features are then passed to a genetic algorithm which is then fitted to a conventional classifier. In 2017, Geoffrey Hinton *et al.* [22] proposed Capsule Networks based on traditional CNN. Based on it, Jaiswal *et al.* [23] introduced Generative Adversarial Capsule Networks (CapsuleGAN), a framework that employs a capsule network as a discriminator in Generative Adversarial Networks (GANs), an innovative approach that has shown excellent performance in modelling image data distributions on the MNIST and CIFAR-10 datasets.

The above work inspired us to investigate an automated micro-expression analysis system for ASD patients.

3. Methodology

3.1. YOLOv8-SMART based dataset

YOLOv8 [24] is an efficient object detection algorithm with a deep learning architecture consisting of multiple components at its core. These components include a backbone network, a feature pyramid network, a prediction head, an anchor frame, and a loss function, which work in concert to make YOLOv8 excel in detection speed and accuracy. YOLOv8 uses a similar backbone as YOLOv5 but replaces all the C3 modules with the more gradient-flow-rich C2f module. The C2f module connects high-level features with contextual information through cross-level connections combination to enhance feature representation [25]. In order to solve the problems associated with detecting close-ups such as facial micro-expressions using the YOLOv8 network, we propose the YOLO-SMART algorithm, as shown in Figure 2. We replaced the C2f module with VOVGSCSP, added MCALayer, and replaced Box Loss with MPDIoU Loss which provides more accurate feedback on positioning accuracy.

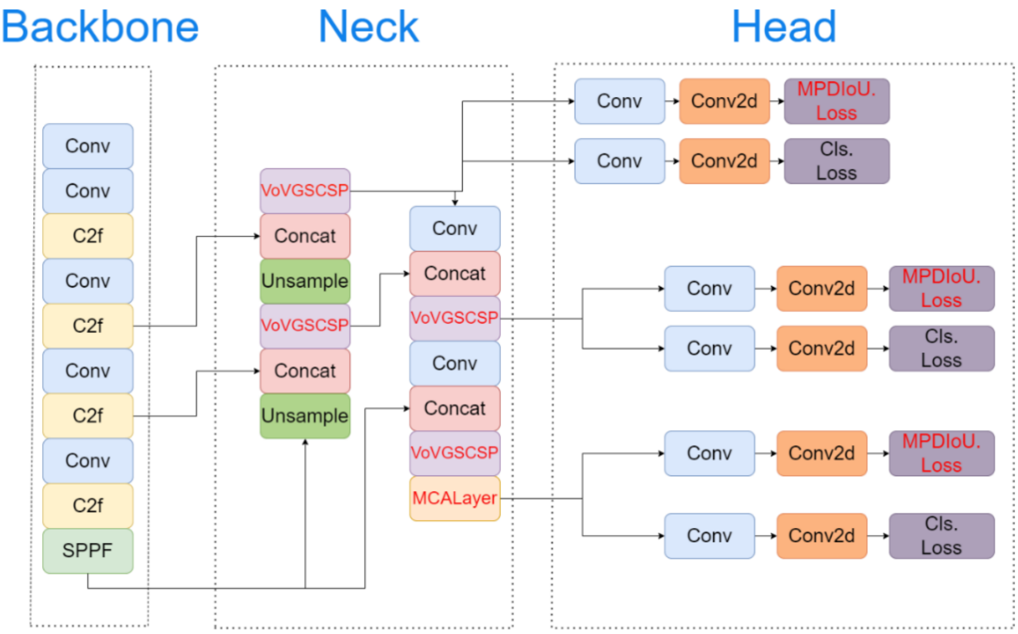


Figure 2. YOLO-SMART architecture diagram.

3.2. VoVGSCSP

We replaced the standard convolution in C2f with the VoVGSCSP module. The multi-scale feature extraction and spatial pyramid structure used in this module allows the model to capture both large-scale and fine detail information. The use of different sized convolution kernels extends the receptive field, which captures a wider range of context and helps to understand the relationship between overall facial features and local details.

As shown in Figure 3, Conv layers are used to extract base features. The first GSConv layer introduces sparsity constraints to optimise feature selection and improve the learning efficiency of the model. The second GSConv layer further refines the features to capture more complex patterns. After that, the Concat layer merges feature maps from different sources to form a rich and comprehensive feature representation. Finally, Conv processes the spliced features to improve the feature representation.

In summary, the Neck part of the VoVGSCSP effectively captures complex facial expression variations through sparsity constraints and multi-layer feature extraction, leading to more accurate micro-expression analysis and classification.

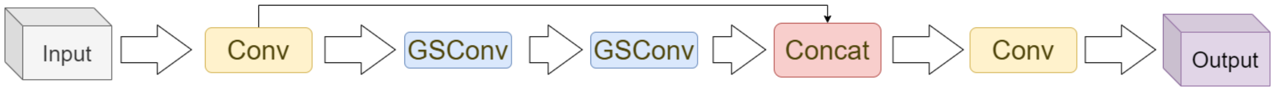


Figure 3. The architecture of VoVGSCSP.

3.3. MCALayer

The core function of MCALayer is feature weighting, which determines the significance of features by adaptively calculating attention weights for each channel. This allows the model to emphasize important features while suppressing redundant or irrelevant ones, thereby enhancing robustness against noise and disturbances. For example, when background lighting changes in an image, MCALayer can boost relevant features, leading to improved accuracy in micro-expression detection. Additionally, MCALayer effectively integrates features from different channels, strengthening the complementary aspects of multimodal data. The architecture diagram of MCALayer is illustrated in Figure 4.

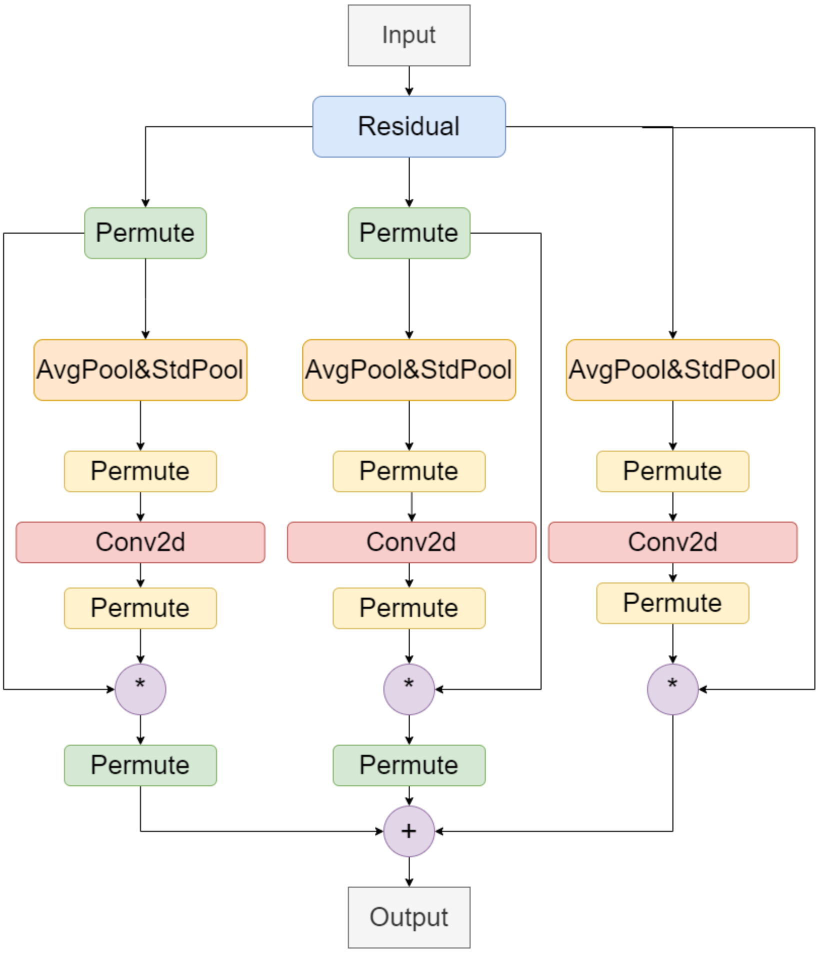


Figure 4. The structure of MCALayer.

In order to verify the effectiveness of MCALayer, we compare it with existing attention mechanism modules SE blocks, CBAM and ECA to evaluate its performance in several key performance indicators.

As observed from the results presented in the table, MCALayer demonstrates a significant improvement in accuracy compared to other modules, especially when compared to the CBAM block. Although MCALayer does not exhibit a clear advantage in terms of training time and memory consumption, the notable enhancement in accuracy compensates for these limitations. This highlights MCALayer's considerable strengths in feature learning and model performance, underscoring its potential for tasks where precision is paramount. Therefore, considering all factors, MCALayer is selected as the attention mechanism module.

3.4. MPDIoU loss

In the YOLO-v8 model, IoU is often used as part of box loss and is widely regarded as a way to calculate regression loss. It is used to assess the degree of overlap between the predicted and true frames. When the degree of overlap is higher, it represents a better effect.

$$\mathbf{IoU} = \frac{\mathbf{area}(\mathbf{truth\ box} \cap \mathbf{pred\ box})}{\mathbf{area}(\mathbf{truth\ box}\ \cup \mathbf{pred\ box})} \tag{1}$$

By minimizing the IoU in Eq. (1), the prediction of the bounding box can be significantly optimized, thus improving the accuracy of the target location. Although IoU is widely used and has good performance in target detection tasks, it is still unable to accurately capture small targets and deal with boundary blurring. In view of this, we introduce MPDloU Loss [26].

MPDloU Loss is a new bounding box similarity comparison metric based on the minimum point distance, which focuses on the specific positional differences of the boxes by minimizing the Euclidean distances between the top left and bottom right points between the predicted bounding box and the actual labelled box.

$$\mathbf{d_1^2 = (x_1^a - x_1^b)^2 + (y_1^a - y_1^b)^2} \tag{2}$$

$$\mathbf{d_2^2 = (x_2^a - x_2^b)^2 + (y_2^a - y_2^b)^2} \tag{3}$$

$(\boldsymbol{x_1^a}, \boldsymbol{y_1^a})$: coordinates of the point in the upper left corner of the prediction box.
$(\boldsymbol{x_2^a}, \boldsymbol{y_2^a})$: coordinates of the point in the bottom right corner of the prediction box.
$(\boldsymbol{x_1^b}, \boldsymbol{y_1^b})$: coordinates of the point in the upper left corner of the actual box.
$(\boldsymbol{x_2^b}, \boldsymbol{y_2^b})$: coordinates of the point in the upper left corner of the actual box.

$$\mathbf{MPDIoU = IoU - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2}} \tag{4}$$

$\boldsymbol{w}$ refers to the width of the image and $\boldsymbol{h}$ refers to the height of the image.

MPDloU Loss focuses on edge pixels and gives them a higher weight in the loss calculation. This approach makes the model focus more on edge information during the training process, which enhances the sensitivity to the target contour and the accuracy of edge detection.

4. Experiment and analysis

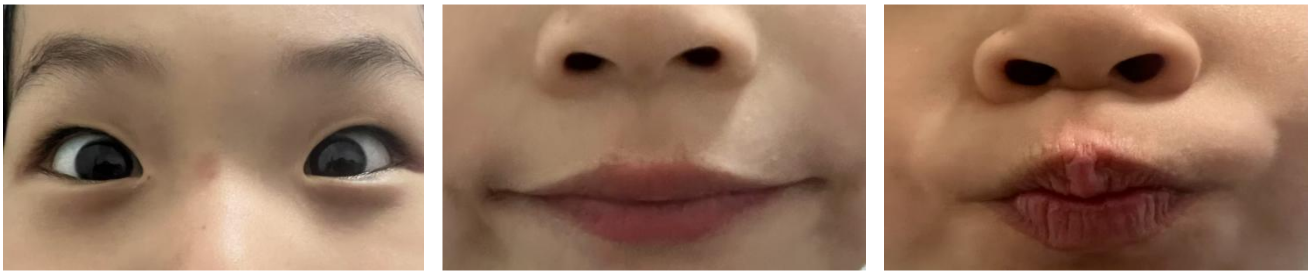
4.1. Data setting

In this study, a movie role-based video dataset [27] for ASD micro-expression detection is created. The dataset is drawn from 10 carefully screened and processed movies. The characters in these movies vividly portrayed people with ASD, thus covering a wide range of situations that people with ASD may face and presenting complex and realistic emotional expressions. To ensure the broad applicability of the dataset, special attention was given to age, gender, and ethnic diversity. The personas in the dataset span an age range of 5 to 40 years, encompassing children, adolescents, and adults, with a balanced gender ratio of 1:1. The dataset includes characters from diverse ethnic backgrounds, representing Caucasian, Asian. This diverse population design ensures that the micro-expression detection model demonstrates strong generalization ability across individuals of varying ages, genders, and ethnicities.

To conduct a thorough analysis of the micro-expressions of characters and their changes throughout the movie, we utilize video editing tools to accurately capture clear footage of individuals with ASD. Given that micro-expressions can last as briefly as 3 milliseconds, we divided them into frames using a frame rate of 24 frames per second, resulting in a sequence of consecutive camera frames. As a result, we created a dataset comprising 972 video clips, each lasting between 2 and 10 seconds, showcasing the emotional expressions of the characters.

To create a comprehensive dataset, manual annotation is employed for the micro-expression labeled image frames. This approach ensures high accuracy and sensitivity in capturing the diversity of micro-expressions. A macro-expression plus micro-expression annotation method is utilized, based on the Facial Action Coding System [28] developed by Paul Ekman and Wallace Friesen. This framework facilitates the establishment of a standardized classification of micro-expressions, comprising 42 Action Units (AUs), as illustrated in Figure 5.

The classification of emotions is refined using specific combinations of micro-expression AU codes. Seven distinct emotion types are identified, as illustrated in Figure 6. Type 01 corresponds to disgust, characterized by AU9, AU15, and AU16. Type 02 represents anger, comprising AU4, AU7, and AU21. Type 03 indicates fear, which includes AU1, AU2, AU4, AU5, AU7, AU20, and AU26. Type 04 is associated with sadness, featuring AU4, AU6, and AU15. Type 05 reflects happiness, defined by AU6 and AU12. Type 06 signifies contempt, consisting of AU12 and AU14. Finally, Type 07 denotes surprise, characterized by AU2, AU5, and AU26.



(a) AU6 **(b)** AU15 **(c)** AU23

Figure 5. Three types of Action Units (AU): **(a)** Action Units (AU) 6: Cheek Raiser; **(b)** Action Units (AU) 15: Lip Corner Depressor; **(c)** Action Units (AU) 23: Lip Tightener.

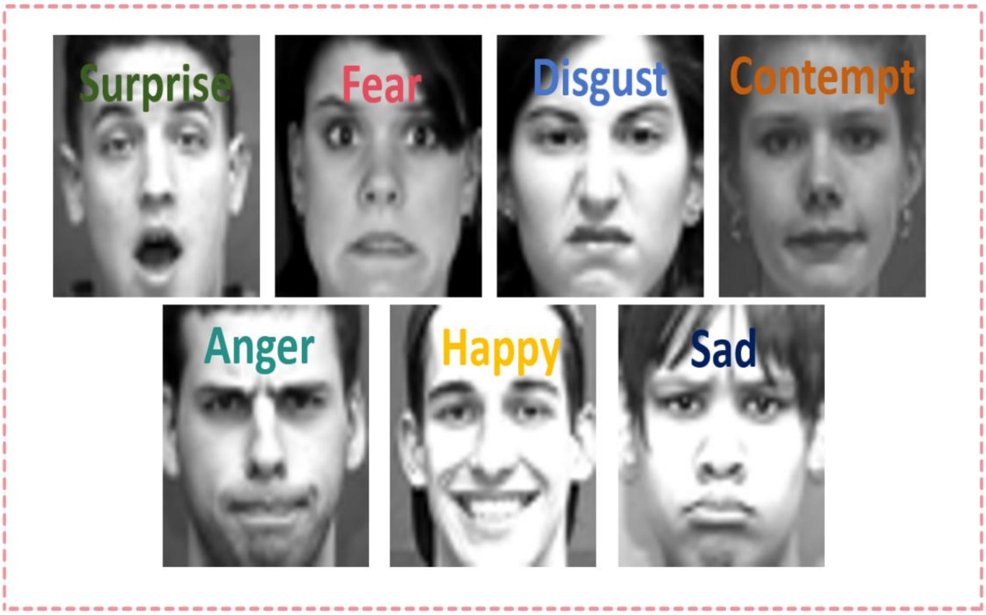


Figure 6. Seven emotion types [29].

In the labeling process, micro-expression images are tagged using the format 00X-0X-000X-000X-0XXX-XX-XX-XX. In this format, 00X-0X-000X-000X indicates the source of the image frame; for example, 003-06-0007-0008 corresponds to the eighth frame of the seventh shot of the sixth character in the third movie. The segment 0X represents the emotion type, while XX-XX-XX denotes specific AU codes. This dataset was ultimately constructed by five biomedical engineers through cross-validation, resulting in a total of 2,944 micro-expression images.

4.2. Evaluation metrics

This study compares the YOLOv8-SMART algorithm with the original object detection algorithm using key evaluation metrics, including Precision, Recall, F1 Score, and Mean Average Precision (MAP). These metrics are crucial for assessing the performance of object detection algorithms. Precision measures the accuracy of positive predictions, while Recall evaluates the algorithm's ability to identify all relevant instances. The F1 Score serves as a harmonic mean of Precision and Recall, providing a balanced assessment of performance. MAP summarizes precision across various thresholds, offering insights into the model's ranking capabilities. To ensure the robustness of our analysis, repeated experiments were conducted, and statistical measures such as mean and variance were employed to evaluate the consistency of the results. This comparative framework effectively highlights the differences between the improved YOLO-V8 algorithm and the original object detection algorithm.

4.3. Results and evaluation

4.3.1. Main experiment

In this experiment, we use a computing platform with NVIDIA RTX 3060 GPUs, 32 GB of RAM, with the PyTorch deep learning framework (version 1.9.0), as well as libraries such as OpenCV for data processing and visualization of results. The dataset is divided into training set, validation set and test set in the ratio of 8:1:1. 100 training rounds are set up, during which we regularly monitor training losses and validation metrics to adjust hyperparameters in a timely manner.

In the testing phase, Mean Average Precision (MAP) and F1-score are utilized as the primary evaluation metrics to comprehensively assess the model's performance. MAP effectively reflects the average detection accuracy across various categories, offering a clear overview of overall performance, which is particularly crucial in multi-class target detection tasks. By calculating precision and recall for each category and averaging them across different thresholds, MAP illustrates the capability of model to recognize diverse micro-expressions. Meanwhile, the F1-score integrates precision and recall, enhancing its effectiveness in addressing unbalanced datasets. By combining these two metrics, we can evaluate the effectiveness of model in the micro-expression recognition task in a more holistic manner, providing a solid foundation for subsequent improvements and optimizations.

The F1 score of YOLOv8-SMART in the experiments increases with the confidence level and reaches 98.9% at a confidence level of 1.0. Figure 7 shows the confusion matrix for detecting micro-expression image data on the test set using YOLOv8-SMART. The analysis of the confusion matrix reveals that the AU combination categories 405-495 and 705-740 show poor classification performance. Category 405-495 pertains to AU57 (head forward) and AU58 (head backward), while category 705-740 involves AU6 and AU7 (eye movements). The system struggles to detect these subtle movements, as they are low-intensity AUs.

To address this issue, a continuous time threshold detection method was applied for low-intensity AUs. When the duration of the movement is less than 200ms, it is classified as a low-intensity AU. If the duration exceeds 200ms, it is categorized as a high-intensity AU.

This method allows for better differentiation between short-duration, subtle movements and more pronounced facial expressions, ultimately improving the system's overall classification accuracy, particularly for low-intensity AUs.

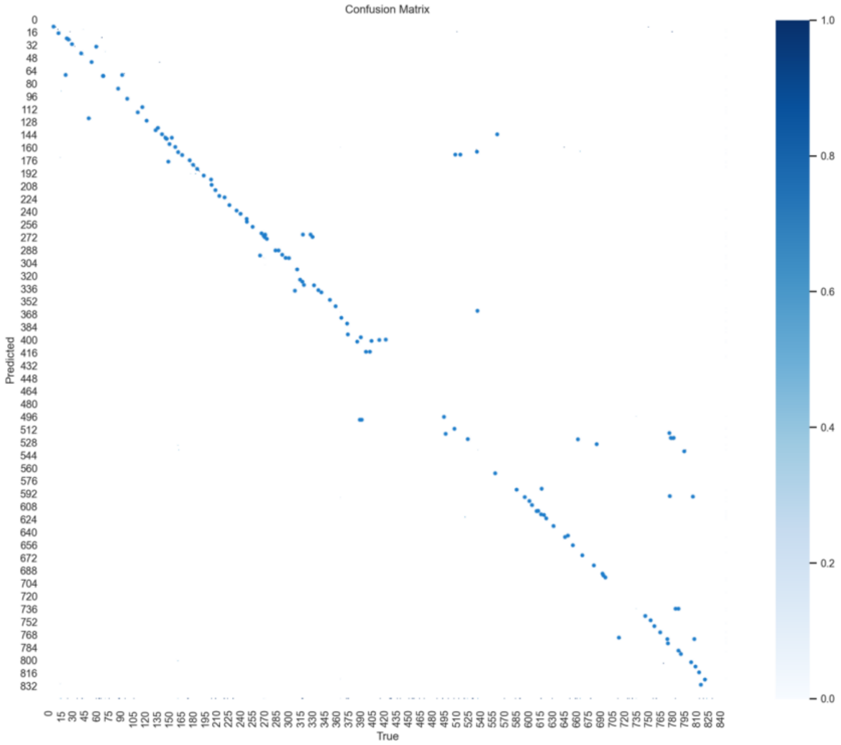


Figure 7. Confusion matrix of YOLOv8-SMART.

4.3.2. Comparative experiment

In previous experiments, we evaluated the performance of the YOLOv8-SMART model for micro-expression detection in ASD patients, and next we compared it with five other target detection algorithms. These five target detection algorithms are YOLO-v3, YOLO-v5, YOLO-v8, Faster R-CNN and SSD.

YOLO. YOLO neural network treats the target detection task as a regression problem and combines various processes such as candidate box extraction, feature extraction and target classification to achieve fast detection. In the YOLO architecture, the input image is divided into an S × S grid, where each grid cell is responsible for predicting several candidate bounding boxes. These boxes are uniformly distributed throughout the image to provide comprehensive coverage of potential objects. Each candidate box has a set of parameters including its position in the grid and a confidence score.

$$\mathbf{confidence} = \mathbf{pr}(\mathbf{object}) * \mathbf{IoU} \tag{5}$$

The confidence score indicates the likelihood that the box contains an object and is calculated based on the model's prediction of the object class and the accuracy of the bounding box coordinates. The algorithms in the YOLO series detect objects based on features throughout the image, rather than multiple detections in specific areas. This approach offers significant advantages in terms of speed and allows real-time processing.

Faster R-CNN. Faster R-CNN is a method built on fast region convolutional networks. Its feature extraction network employs a VGG architecture with 13 convolutional layers and 4 pooling layers. Unlike traditional R-CNN, Faster R-CNN introduces a Region Proposal Network (RPN) that efficiently generates candidate regions likely to contain targets. These regions are then passed to the following networks for target classification and bounding box regression. The accompanying figure illustrates the architecture of Faster R-CNN.

SSD. SSD is a classic network in single-stage object detection algorithms. It consists of three main parts: the backbone network, feature extraction network, and detection network. VGG16 is used as the base model, with an additional convolutional layer added to generate more feature maps, enhancing the network's ability to detect targets in images of varying sizes for micro-expressions. In the detection network section, a fixed number of prior boxes are generated at every pixel location of each feature map, with the size of the prior boxes determined by the scale S of the input image. Then, the detection network generates these fixed numbers of prior boxes at all pixel points of each feature map.

$$\mathbf{S} = \mathbf{s}_{\mathbf{min}} + \frac{\mathbf{s}_{\mathbf{max}} - \mathbf{s}_{\mathbf{min}}}{\mathbf{m} - \mathbf{1}}(\mathbf{x} - \mathbf{1}) \tag{6}$$

where $\boldsymbol{x}$ denotes the feature map, $\boldsymbol{s_{min}}$ denotes the minimum value of the size scale, and $\boldsymbol{s_{max}}$ denotes the maximum value of the size scale.

In this experiment, the dataset is divided into training, validation and test sets in the ratio of 8:1:1. We use a dataset based on micro-expression images from ASD movies to train the CNN architecture. MAP is one of the most important performance metrics of the model. In model training, YOLOv8-SMART has the highest MAP. Table 1 reports the precision, recall, F1 score and MAP of the six target detection algorithms in the micro-expression detection experiments in this dataset.

To better visualize the improvements of YOLOv8-SMART, we conduct the comparison experiments three times and calculate the average values of each metric in Table 2. Additionally, multiple sets of histograms Figure 8 are used to compare the average values of evaluation metrics for different algorithms.

Table 1. Performance comparison of attention mechanism modules.

	Accuracy(%)	Training Time(h)	Memory Consumption(MB)
MCALayer	94.78	14.3	6.382
SE	89.2	13.8	5.248
CBAM	80.3	16.2	7.281
ECA	87.4	14.1	5.10

Table 2. Evaluation of six target detection algorithms on various metrics.

Algorithm	Precision	Recall	F1 Score	MAP
YOLO-v3 (1)	0.983	0.71	0.30	0.550
YOLO-v3 (2)	0.987	0.70	0.35	0.543
YOLO-v3 (3)	0.976	0.69	0.31	0.551
YOLO-v5 (1)	**0.993**	0.71	0.25	0.501
YOLO-v5 (2)	0.990	0.65	0.27	0.489
YOLO-v5 (3)	0.991	0.70	0.25	0.492
YOLO-v8 (1)	0.77	0.70	0.32	0.538
YOLO-v8 (2)	0.74	0.68	0.33	0.540
YOLO-v8 (3)	0.77	0.69	0.31	0.521
YOLOv8-SMART (1)	0.989	0.72	0.35	**0.571**
YOLOv8-SMART (2)	0.983	0.71	0.33	0.569
YOLOv8-SMART(3)	0.988	**0.73**	0.36	0.566
Faster R-CNN (1)	0.70	0.67	0.57	0.552
Faster R-CNN (2)	0.72	0.67	0.59	0.554
Faster R-CNN (3)	0.68	0.69	0.61	0.547
SSD (1)	0.80	0.64	**0.66**	0.453
SSD (2)	0.81	0.60	0.64	0.446
SSD (3)	0.78	0.65	0.67	0.451

Through comparative experiments, YOLOv8-SMART has better evaluation metrics parameters and therefore has the best target detection. The average precision of the six target detection algorithms is 0.982 (YOLO-v3), 0.991 (YOLO-v5), 0.76 (YOLO-v8), 0.987 (YOLOv8-SMART), 0.70 (Faster R-CNN), and 0.797 (SSD). The recall of the YOLO family of algorithms is much higher than that of Faster R-CNN and SSD. YOLOv8-SMART has the highest average MAP of 0.568.

Therefore, the improved YOLOv8-SMART algorithm shows significant effectiveness in the field of target detection, especially in micro-expression recognition.

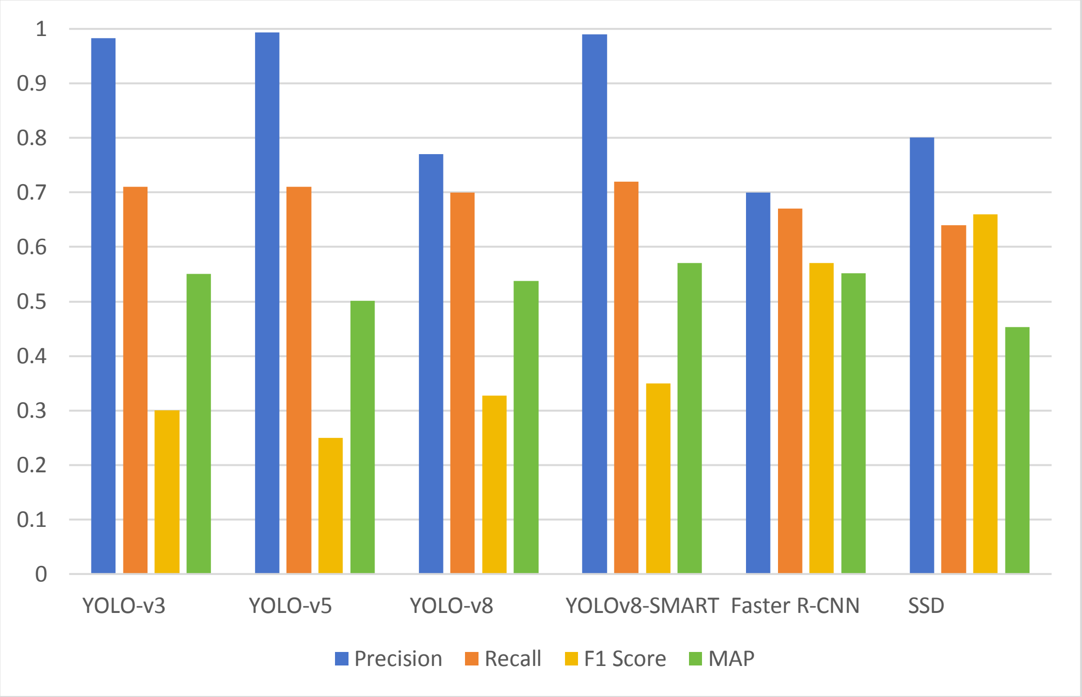


Figure 8. Histogram of average evaluation metrics.

4.3.3. Random repeated experiment

Random repeated of the experiment not only helps to eliminate the effects of chance but also identifies fluctuations in the performance under various conditions. We conduct five random repeated experiments, each involving a random division of the dataset into training, test, and validation sets. For each experiment, the evaluation metric MAP Is recorded to measure the detection performance of the model. The results indicate that the mean MAP across the five experiments is 0.570, with a standard deviation of 0.00856. This suggests that the performance of the model remains relatively stable throughout the experiments. To present the results more intuitively, we plot Figure 9, which clearly illustrates the distribution of MAP values across the experiments.

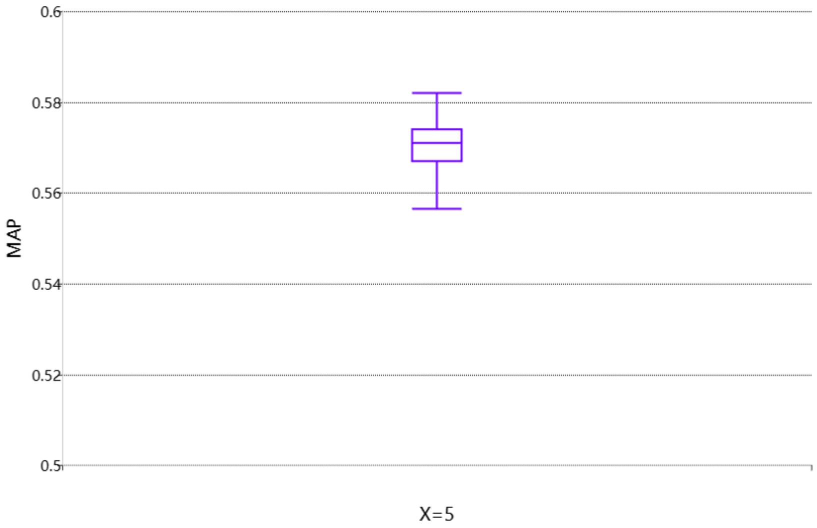


Figure 9. Boxplot of five random repeated experiments.

4.4. Discussion and limitations

At the current stage, pharmacologic treatment of ASD focuses primarily on alleviating the symptoms associated with the disorder rather than curing autism. Available evidence suggests that some of these medications have significant side effects in children [30]. Non-pharmacological treatments that show solid evidence of improving verbal communication in individuals with ASD include music therapy, social behavioral therapy [31].

With the development of digital technology, new analytical tools and methods have been introduced to the field of film studies. Cinemetrics, as an emerging research paradigm, measures and compares stylistic differences in films through a quantitative framework. When applied to ASD-themed films, Cinemetrics helps to reveal the subtleties of character micro-expressions, providing viewers with a deeper emotional understanding.

Automated micro-expression recognition algorithms developed for individuals with ASD have made significant strides in interpreting emotional expressions within this population. The ability to accurately detect and analyze micro-expressions can enhance communication strategies and aid in emotion recognition, which often poses challenges for those with ASD. Our algorithm demonstrates promising results in identifying key micro-expressions. By offering objective insights into emotional responses, this technology has the potential to improve interactions and foster better understanding between individuals with ASD and their peers or caregivers.

The current dataset encompasses a wide range of camera positions and angles, including various distances, low and high angles, and diverse lighting conditions. During the labeling process, we incorporated AU51-AU58, which are specifically designed to capture head posture and movement, thereby enriching the information available for facial expression recognition. Despite these efforts, practical applications may still encounter challenges, such as extreme lighting conditions (e.g., strong backlighting or overexposure) and partial occlusion (e.g., from hair, glasses, or hands), which can lead to the loss or misidentification of facial keypoints. Additionally, in some cases, faces in near and middle views may appear disproportionately large, while detection frames in distant views may be too small, compromising the algorithm's ability to accurately recognize micro-expressions and ultimately reducing detection accuracy.

These challenges can impact the stability and reliability of facial expression analysis systems in real-world settings, such as telemedicine and mental health assessments. Nevertheless, this study provides a novel perspective on micro-expression recognition, highlighting the need for further research to enhance both the comprehensiveness and practical applicability of such systems. Future work will focus on improving the model's robustness in complex environments and enhancing its ability to adapt to varying conditions across different clinical settings.

5. Conclusion and future work

The micro-expressions displayed by individuals with ASD provide valuable insights into their emotional state. By utilizing video frame-splitting techniques, we can capture these fleeting expressions to gain a more nuanced understanding of the emotional experience of people with ASD. This approach has great potential for healthcare professionals seeking to more accurately characterize the psychological state of people with ASD.

In future studies, our research aims to extend the current dataset by increasing the number of micro-expression images to include images of ASD patients of all ages from real life. This extension is essential to enhance the robustness and applicability of our micro-expression recognition system. We are also committed to exploring and improving various target detection algorithms to evaluate their effectiveness in recognizing micro-expressions and to determine the most effective methods for ASD micro-expression analysis.

In addition, we plan to collaborate with experts in the field of ASD to gain a deeper understanding of the unique challenges and nuances of this disorder. This collaborative approach will allow us to improve our models and algorithms to better meet the specific needs of individuals with ASD, ultimately leading to more effective diagnostic tools and therapeutic interventions.

Acknowledgments

This work is supported by the "National Natural Science Foundation of China" (No. 82220108007). Chen Li and Ning Xu have equal contributions as corresponding authors in this paper. Since our dataset involves copyright issues, the example image uses an open source movie dataset.

Authors' Contribution

Yutong Gu: data curation, methodology, experiments, evaluation, writing—review and editing; Hanni Li: data curation, experiments, writing; Jiarong Liu: data curation; Chenxi Liu: data curation; Yuxuan Li: experiments; Chen Li: data curation, methodology, evaluation, writing—review and editing, funding acquisition; Ning Xu: data curation, writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

References

[1] Geschwind DH. Genetics of autism spectrum disorders. *Trends Cognit. Sci.* 2011, 15(9):409–416.

[2] Prevalence of autism spectrum disorders-Autism and developmental disabilities monitoring network. *MMWR Surveill Summ* 2009, 58:1–20.

[3] Baird G, Simonoff E, Pickles A, Chandler S, Loucas T, *et al*. Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). *The Lancet* 2006, 368(9531):210–215.

[4] Autism. World Health Organization. 2023. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (accessed on 5 October 2024).

[5] Eissa N, Al-Houqani M, Sadeq A, Ojha SK, Sasse A, *et al*. Current enlightenment about etiology and pharmacological treatment of Autism Spectrum Disorder. *Front. Neurosci.* 2018, 12:304.

[6] Li X, Pfister T, Huang X, Zhao G, Pietikäinen M. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, April 22–26, 2013, pp. 1–6.

[7] Ekman P, Friesen WV. Nonverbal leakage and clues to deception. *Psychiatry*, 1969, 32(1): 88–106.

[8] Quang NV, Chun J, Tokuyama T. CapsuleNet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, May 14-18, 2019, pp. 1–7.

[9] Tsivian Y. *Digital Tools in Media Studies*; Michael Ross, Manfred Grauer, Eds. Cinemetrics, part of the humanities' cyberinfrastructure; Bielefeld: Transcript Verlag, 2009.

[10] Jung JJ, You E, Park SB. Emotion-based character clustering for managing story-based contents: A cinemetric analysis. *Multimedia Tools Appl.* 2013, 65(1):29–45.

[11] Internet Archive. Movies and Videos. Internet Archive. Available: https://archive.org/details/movies. (accessed on 10 October 2024).

[12] Genovese A, Butler MG. Clinical assessment, genetics, and treatment approaches in Autism Spectrum Disorder (ASD). *Int. J. Mol. Sci.* 2020, 21(13):4726.

[13] Kanner L. Autistic disturbances of affective contact. *Nervous Child* 1943, 32:217–253.

[14] Mezzacappa A, Lasica P-A, Gianfagna F, Cazas O, Hardy P, *et al*. Risk for Autism Spectrum Disorders according to period of prenatal antidepressant exposure: A systematic review and meta-analysis. *JAMA Pediatrics* 2017, 171(6):555–563.

[15] American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5*. 5th ed., American Psychiatric Publishing, 2013.

[16] Quang NV, Chun J, Tokuyama T. CapsuleNet for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, May 14–18, 2019, pp. 1–7.

[17] Haggard EA, Isaacs KS. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of Research in Psychotherapy*, Boston: Springer US, 1966, pp. 154–165.

[18] Ekman P, Friesen WV. Facial action coding system: A technique for the measurement of facial movement. *Palo Alto*. 1978(3).

[19] Li X, Pfister T, Huang X, Zhao G, Pietikäinen M. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, April 22–26, 2013, pp. 1–6.

[20] Li X, Hong X, Moilanen A, Huang X, Pfister T, *et al*. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Trans. Affective Comput.* 2018, 9(4):563–577.

[21] Devangini P, Hong X, Zhao G. Selective deep features for micro-expression recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, December 4–8, 2016, pp. 2258–2263.

[22] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. *Advances in neural information processing systems* 2017, 30.

[23] Jaiswal A, AbdAlmageed W, Wu Y, Natarajan P. CapsuleGAN: Generative adversarial capsule network. In *Proceedings of the European conference on computer vision (ECCV) workshops*, Munich, Germany, September 8–14, 2018, pp.1586–1601.

[24] Ultralytics. Ultralytics. Github. 2024. Available: https://github.com/ultralytics/ultralytics (accessed on 1 October 2024).

[25]Wu T, Dong Y. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. *Applied Sciences*. 2023, 13(24):12977.

[26] Ma Siliang, Xu Yong. MPDIoU: A loss for efficient and accurate bounding box regression. *arXiv* 2023, arXiv: 2307.07662.

[27] Yutong Gu, Hanni Li, Yuxuan Li, Jiarong Liu, Chenxi Liu, *et al*. A movie role based video dataset for Autism Spectrum Disorder micro-expression detection. In *Proceedings of Conference on Machine Vision, Image Processing and Imaging Technology (MVIPIT 2024)*, Zhangjiakou, China, September 13–15, 2024.

[28] Ekman P, Friesen WV. Facial action coding system. *APA PsycTests*. 1978.

[29] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, *et al*. The *Extended Cohn-Kanade Dataset (CK+)*. Pittsburgh: Carnegie Mellon University, 2010. Available: http://www.pitt.edu/~emotion/ck-spread.htm (accessed on 10 October 2024).

[30] McPheeters ML, Warren Z, Sathe N, Bruzek JL, Krishnaswami S, *et al*. A systematic review of medical treatments for children with Autism Spectrum Disorders. *Pediatrics* 2011, 127(5):e1312–e1321.

[31] Samata SR, Gonda X, Tarazi FI. Autism Spectrum Disorder: Classification, diagnosis and therapy. *Pharmacol. Ther.* 2018, 190:91–104.