ABSTRACT Modelling the visual attention of people with autism spectrum disorder (ASD) is attracting more and more interest. This consists in determining where ASD people look and in inferring the important visual features contributing to the gaze deployment. In this article, we investigate whether or not existing neurotypical as well as ASD saliency models perform well over the whole spectrum of autism. For this purpose, we propose two new eye-tracking datasets of ASD people in order to cover a large part of the autism spectrum, going from high-level functioning (e.g. Asperger) to low-level functioning (e.g. Kanner) autism. We demonstrate that current neurotypical and ASD models do not generalize well and perform well only on a small part of the spectrum. Our objective is to raise the awareness of computer scientists to the difficult task we are facing up when it comes to simulate the gaze deployment of ASD people.

INDEX TERMS Visual attention, eye movements, saliency model, autism, ASD.

I. INTRODUCTION

Autism is a psychic structure with a specific mind inducing a singular relation to the world. It affects 1 in 59 children [1]. Autism spectrum disorders (ASD) encompass a number and various syndromes, going from Asperger to Kanner syndromes. They consist of impairments in social interaction (such as voice retention and gaze withdrawing), communication and repetitive pattern of behaviors. These syndromes lie on a spectrum which reflects a continuum in the severity of troubles.

Social visual engagement difficulties are one of the most noticeable sign of autism. Many studies have therefore investigated the use of eye-based methods to understand and investigate autism syndromes (see [2] for a review of such methods). Measuring eye movements thanks to eye-tracking devices is now very common and rather easy. A wealth of information can be extracted from the collected eye-tracking data, such as fixation locations, durations to name a very few [3]–[5].

There is currently a growing interest in determining the visual factors that attract or repel the visual attention of ASD people. Thanks to new eye-tracking datasets involving ASD people [6], the number of saliency models is significantly increasing [7]–[9]. All these models aim to output a 2D saliency map from an input image. The saliency map indicates where ASD people look at. Predicting saliency maps is important for several reasons. First, accurate saliency maps can boost the accuracy of ASD diagnosis by better distinguishing people with ASD from neurotypical people as proposed in [10]. Second, by training deep neural network with neurotypical and ASD eye-tracking data, it is possible to visualize the learned (or deep) features and to interpret their differences and influences in the visual deployment of ASD people. For instance, in [11], authors showed that ASD people have a significantly greater image center, background, and pixel-level bias, but a reduced object-level bias and semantic-level bias.

As indicated previously, the most recent and performant saliency models rely on deep-based architectures [7]–[9], [12]. The training procedure is generally composed of two steps in order to deal with the lack of eye-tracking data of...
ASD people. The first step consists in training the model on eye-tracking dataset of neurotypical observers, for which a number of datasets exist [13]–[16]. The second step consists in fine-tuning the model on the ASD eye-tracking dataset for a more effective saliency prediction.

According to reported performances, these models perform well in predicting where ASD people look at. However, this conclusion raises a number of concerns and hides the complexity of the task we are facing up.

In this article, we want to increase the awareness of computer scientists on the lack of generalization of the proposed models. For that purpose, we propose two new eye-tracking datasets involving ASD people. These datasets are very different from an existing one [6], since the degree of autism of subjects is much more severe. Thanks to these 3 datasets, we cover a large part of the autism spectrum and we are able to test the ability of existing saliency models to predict where ASD people look at. As expected, we show that there is a high discrepancy in the conclusion due to the differences within the populations of ASD people.

The paper is organized as follows. In section II, we present the three eye-tracking experiments involving ASD people. Section III presents the performances of neurotypical saliency models as well as one ASD model trained with eye-tracking data collected on ASD people. The last section IV discusses the results.

II. EXPERIMENTAL STUDIES

In this section, we present the three eye-tracking datasets used in this study. The former is the dataset proposed for the SaliencyForASD challenge which took place during the International Conference on Multimedia and Expo (IEEE ICME) 2019. This dataset, called in the following ICME, is presented in section II-A. Section II-B presents a new eye-tracking dataset involving two sets of ASD people. We will call them MIE Fo and MIE No.

A. ICME DATASET

The ICME dataset is described in [6]. In the following subsections, we remind the key elements of this dataset. Table 1 also provides a quick overview of its main features. The dataset can be accessed in http://doi.org/10.5281/zenodo.2647418

| Category      | Items                          | Details       |
|---------------|-------------------------------|---------------|
| Stimuli       | Image number                  | 300           |
| Eye tracker   | Model                         | Tobii T120    |
|               | Size                          | 17 inch       |
|               | Resolution                    | 1280 x 1024   |
|               | Sampling rate                 | 120Hz         |
|               | Tracking distance             | 65cm          |
| Subjects      | ASD subjects number           | 14 reserved   |
|               | Mean age                      | 6 discarded   |
|               |                               | 8 yo          |
| Experimental  | View time                     | 3s            |
| protocol      | Gray interval                 | 1s            |
|               | Presentation order            | Random        |
|               | Session number                | 10            |
|               | Task condition                | Free-viewing  |

1) PARTICIPANTS

Twenty high-functioning children with ASD were involved in the experiment. Data from six people were discarded because of the poor quality of eye tracking data. Remaining people are from 5 to 12 years old, with an average of 8 years old. The sample is diagnosed autistic according to DSM-V criteria [17].

2) APPARATUS

A Tobii T120 Eye Tracker was used to record eye movements, with a sampling frequency of 120Hz. Images were displayed on a 17 inches screen with a resolution of 1280 × 1024, at a distance about 65 cm from the eye tracker.

3) STIMULI

Three hundred stimuli were used. They were chosen in the eye tracking dataset proposed in [14]. Figure 1 presents 8 samples of the images used during the test.

These images were shuffled into 10 sequences of 30 images, in order to limit the eye tracking experiment duration. Images were displayed during 3 seconds, followed by a one-second gray background. Calibration was carried out at the beginning of each sequence. Before the experiment, all subjects were told to look at stimuli freely.

B. MIE FO AND MIE NO EYE TRACKING DATA

In this article, we introduce a new eye tracking dataset involving ASD people from two different French Medical
TABLE 2. Presentation of the proposed eye tracking dataset involving two sets of ASD people.

| Category            | Items                      | Details                  |
|---------------------|----------------------------|--------------------------|
| Stimuli             | Image number               | 25                       |
| Eye tracker         | Model                      | SMI RED 500              |
|                     | Size                       | 22 inch x 19.4 inch      |
|                     | Resolution                 | 1920 x 1200              |
|                     | Sampling rate              | 500Hz                    |
|                     | Tracking distance          | 60-70cm                  |
| Subjects            | MIE Fo                     | 17 (13 ± 4°, 4°)         |
|                     | Mean age                   | 16 ± 2 yd                |
|                     | MIE No                     | 12 (9°, 3°)              |
|                     | Mean age                   | 29 ± 7 yd                |
| Experimental        | Viewing time               | 4s                       |
| protocol            | Gray interval              | 1s                       |
|                     | Presentation order         | Random                   |
|                     | Task condition             | Free-viewing             |

Educational Institutes (MIE). We present the details of the proposed experiment in Table 2.

1) PARTICIPANTS
To collect data, we asked to twenty nine ASD people through two French MIEs, namely MIE Fo and MIE No, to participate to the experiment. Seventeen participants come from MIE Fo, with a mean age of 16-year olds (std = 2). The remaining twelve participants come from MIE No and have a mean age of 29-year olds (std = 7).

All participants are mostly Kanner autists [18], meaning that they have elementary defenses, with no access (or very limited) to language and for whom the presence of other people may be very worrying and who require clear and well-defined daily routines. ASD people from MIE Fo have a stronger attraction for knowledge, for learnings and for any opening to the world than MIE No. The autistic troubles of ASD people from MIE No are much more severe than ASD people from MIE Fo, and therefore they manifest a stronger defense against stimuli coming from the outside.

2) APPARATUS
A SMI red 500 remote eye-tracker was used to record participants’ gaze. The sampling frequency is 500Hz. This eye tracker estimates the location of gaze with high accuracy (precision < 1°) based on the reflection of near-infrared light from the cornea and the pupil. It is important to underline that we do not use a chin-rest to guarantee the comfort of participants and to reproduce as much as possible a natural viewing condition.

Stimuli were presented in a full-screen mode on DELL screen U2410 with a screen resolution of 1920 x 1200 pixels. The height and width of the display is 44.26 cm and 55.88 cm, respectively. The participants sat in a comfortable chair at a distance of 60-70 cm from the screen. The stimuli subtended a horizontal visual angle of 49.92° and a vertical visual angle of 43.8°. The number of pixel per degree is then approximately equal to 38 and 27 for the horizontal and vertical dimension, respectively. The experiment has been carried out in a free-viewing task. Participants were asked to look at the screen as naturally as possible.

The raw eye-tracking data is then classified into saccades and fixations thanks to the built-in SMI algorithm. We remind that a saccade is defined as a rapid change in gaze location, and a fixation is regarded as being bordered by two saccades. A velocity-based method is used to determine the saccade event [19].

3) STIMULI
To build the database, 25 stimuli were chosen with a low semantic meaning and a low emotional arousal in order to not disturb subjects. Figure 1 presents samples of the images used during the test.

Images were displayed during 4 seconds, followed by a one-second gray background, within one session. Any specific task has been mentioned, and subjects were accompanied by their tutors so they feel more comfortable and prone to perform the experiment.

C. DISCUSSION
The three sets of eye tracking data, namely ICME, MIE Fo and MIE No, concern people suffering from ASD, meaning they all have problems with social, emotional, and communication skills. However, fundamental differences exist between these three populations with different levels of ability and disability. For the dataset ICME, people are high-functioning, meaning that they need some support for social interactions, organization, etc. Such people are located on one side of autistic spectrum, rather close to the Asperger syndrome. On the other side, we have the participants of the MIE No, who present severe autism troubles. They require a strong support for the daily life. In this case, such people are said Kanner [18]. The participants of MIE Fo are inbetween, even if they are much closer to participants from MIE No than ICME participants.

Therefore, with these three populations, we are on the autism spectrum scale, or on the great continuum [20] going from the Kanner syndrome to the Asperger syndrome. Figure 2 represents the autistic spectrum disorders, the symptoms for the different levels of autism. On the same figure, we place approximately the three eye-tracking datasets that are used in this study. ICME dataset is located at the left-hand side whereas both MIE datasets are on the other side.

The question we want to investigate now is the ability of existing saliency models to predict where ASD people look at. We put to the test saliency models that have been trained with neurotypical eye-tracking data and one saliency model that has been trained with ASD eye-tracking data.

For the sake of reproducible research, these three datasets are all freely available on the following link https://www-percept.irisa.fr/asperger_to_kanner/. This includes original stimuli, human saliency maps and fixation maps.
To carry out the evaluation, we use five similarity metrics that are recommended and classically used in saliency benchmark [21]. They are briefly described below that are recommended and classicaly used in saliency benchmark:

- Correlation Coefficient, $CC \in [-1, 1]$, evaluates the degree of linearity between two saliency maps. $CC = 1$ indicates that there is a perfect linear relationship between the two maps;
- SIM, $SIM \in [0, 1]$, represents the similarity between two saliency map distributions, evaluated through the intersection between histograms of saliency. $SIM = 1$ indicates the highest similarity;
- AUC, $AUC \in [0, 1]$, is the area under the Receiver Operating Characteristics (ROC) curve. We classically use three variants of AUC, namely AUC-J, AUC-B and AUC-S. These metrics measure how well the predicted saliency map of an image predicts the ground truth human fixations on the image. The AUC is determined by plotting the ROC curve thanks to binary thresholdings. The difference between AUC-J and AUC-B relies on how true and false positives are calculated. AUC-S differs from the two previous implementations in the fact that fixation locations taken from other images are also used to account for central bias trend. More details are given in [21]–[23];
- IG, $IG \in [0, +\infty]$, is the information gain which consists in comparing the average log-probability of fixated pixels to a given baseline model. In this study, the baseline line is the human saliency map [24];
- KL, $KL \in [0, +\infty]$, is the Kullback-Leibler divergence between the predicted and the human saliency maps. $KL = 0$ indicates a perfect similarity between the two maps.

Mored details about these metrics can be found in [21]–[25].

### III. PERFORMANCES OF SALIENCY MODELS ON NEUROTYPICAL AND ASD DATASETS

In this section, we evaluate the performances of saliency models by comparing the predicted saliency maps with saliency maps computed from neurotypical and ASD eye-tracking data.

#### A. METHOD

To carry out the evaluation, we use five similarity metrics that are recommended and classically used in saliency benchmark [21]. They are briefly described below:

- Correlation Coefficient, $CC \in [-1, 1]$, evaluates the degree of linearity between two saliency maps. $CC = 1$ indicates that there is a perfect linear relationship between the two maps;
- SIM, $SIM \in [0, 1]$, represents the similarity between two saliency map distributions, evaluated through the intersection between histograms of saliency. $SIM = 1$ indicates the highest similarity;
- AUC, $AUC \in [0, 1]$, is the area under the Receiver Operating Characteristics (ROC) curve. We classically use three variants of AUC, namely AUC-J, AUC-B and AUC-S. These metrics measure how well the predicted saliency map of an image predicts the ground truth human fixations on the image. The AUC is determined by plotting the ROC curve thanks to binary thresholdings. The difference between AUC-J and AUC-B relies on how true and false positives are calculated. AUC-S differs from the two previous implementations in the fact that fixation locations taken from other images are also used to account for central bias trend. More details are given in [21]–[23];
- IG, $IG \in [0, +\infty]$, is the information gain which consists in comparing the average log-probability of fixated pixels to a given baseline model. In this study, the baseline line is the human saliency map [24];
- KL, $KL \in [0, +\infty]$, is the Kullback-Leibler divergence between the predicted and the human saliency maps. $KL = 0$ indicates a perfect similarity between the two maps.

### B. PERFORMANCE ON NEUROTYPICAL EYE TRACKING DATASET

We first assess the performances of 5 neurotypical saliency models on the MIT300 dataset: SAM-ResNet and SAM-VGG [26], SalGan [27], DeepGaze II [28] and MLNET [29]. These models are called neurotypical, since they have all been trained with neurotypical eye-tracking data collected over natural scenes. Our first objective is to provide evidence that such models perform rather well to predict where neurotypical observers look at. The MIT300 is composed of 300 natural indoor and outdoor scenes. Eye-tracking data has been collected while neurotypical observers freely watched these stimuli onscreen.

Table 3 presents the degree of similarity between the ground truth and predicted saliency maps. On MIT300, we observe that the best neurotypical models are SAM-ResNet, SAM-VGG and SalGan. For instance, the $CC$ score of SAM-Resnet is 0.78. MLNET is the least accurate model in the context of this test; in term of $CC$ score, MLNET performs at 0.52.

#### C. PERFORMANCE ON ASD EYE TRACKING DATASET

In this section, we first evaluate the ability of aforementioned neurotypical saliency models to predict where ASD people look at.

We also consider Nebout’s model [8], which is a saliency model dedicated to predict ASD saliency. Figure 4 presents the overall architecture of this model, which is inspired from 3 previous saliency models, namely CASNet model [30], deep gaze network [28] and the multi-level deep network of [29]. The model is a two-stream deep architecture, taking fine-resolution and coarse-resolution images as inputs. To extract features from these images, conv3_pool, conv4_pool and conv5_3 layers of VGG-16 are used as the encoder. Then, extracted features go through a shallow network composed of one convolutional layer, a pyramid of dilated convolution and a second convolutional layer. The two streams are then concatenated into a single stream. Next, features are weighted by a point-wise multiplication with the concatenated maps, and a final convolutional layer returns the predicted saliency map. The training procedure consisted in pre-training the network with MIT1003 dataset. The model was then fine-tuned with the ASD ICME dataset. Horizontal flip was used to augment the data. Also, the model takes in consideration the positional bias of autistic people (i.e. the mean saliency map of the training dataset). More information are given in the paper [8]. In the following subsections,
TABLE 4. Performances over ASD datasets of neurotypical deep models and one dedicated model. Best scores are shown in bold, and standard deviation in brackets.

| Dataset | Model       | CC ↑ | IG ↑ | KL ↑ | NSS ↑ | SIM ↑ | AUC-B ↑ | AUC-J ↑ | AUC-S ↑ |
|---------|-------------|------|------|------|-------|-------|---------|---------|---------|
| ICME    | Nebout      | 0.69 (0.11) | -1.74 (0.88) | 1.02 (0.56) | 1.25 (0.38) | 0.66 (0.05) | 0.77 (0.03) | 0.79 (0.04) | 0.61 (0.08) |
|         | SAM-Resnet  | 0.72 (0.13) | -4.60 (2.50) | 3.17 (1.88) | 1.49 (0.54) | 0.63 (0.07) | 0.74 (0.06) | 0.79 (0.05) | 0.66 (0.07) |
|         | SAM-VGG     | 0.60 (0.14) | -4.76 (2.61) | 3.37 (2.00) | 1.32 (0.58) | 0.55 (0.06) | 0.67 (0.06) | 0.78 (0.05) | 0.79 (0.06) |
|         | Sa/GAN      | 0.68 (0.13) | -2.02 (1.29) | 1.37 (0.96) | 1.41 (0.47) | 0.62 (0.06) | 0.74 (0.05) | 0.79 (0.05) | 0.66 (0.06) |
|         | DeepGazeII  | 0.73 (0.15) | -2.44 (1.67) | 1.63 (1.25) | 1.51 (0.59) | 0.65 (0.08) | 0.73 (0.07) | 0.81 (0.05) | 0.63 (0.08) |
|         | MLNET       | 0.60 (0.17) | -1.95 (1.01) | 1.25 (0.67) | 1.32 (0.66) | 0.59 (0.06) | 0.68 (0.06) | 0.79 (0.06) | 0.62 (0.07) |
| MIE Fo  | Nebout      | 0.66 (0.09) | -2.74 (1.28) | 1.53 (0.86) | 1.41 (0.43) | 0.59 (0.04) | 0.77 (0.04) | 0.79 (0.04) | 0.58 (0.06) |
|         | SAM-Resnet  | 0.63 (0.14) | -2.83 (1.55) | 1.54 (1.00) | 1.33 (0.55) | 0.59 (0.06) | 0.70 (0.08) | 0.79 (0.04) | 0.56 (0.06) |
|         | SAM-VGG     | 0.58 (0.16) | -2.99 (1.26) | 1.67 (0.85) | 1.30 (0.62) | 0.56 (0.06) | 0.68 (0.07) | 0.79 (0.04) | 0.60 (0.05) |
|         | Sa/GAN      | 0.62 (0.14) | -2.23 (0.98) | 1.21 (0.62) | 1.39 (0.54) | 0.58 (0.06) | 0.78 (0.05) | 0.79 (0.05) | 0.62 (0.06) |
|         | DeepGazeII  | 0.63 (0.15) | -2.70 (1.16) | 1.57 (0.81) | 1.46 (0.59) | 0.59 (0.06) | 0.72 (0.05) | 0.80 (0.04) | 0.59 (0.07) |
|         | MLNET       | 0.47 (0.19) | -1.99 (0.66) | 0.99 (0.40) | 1.08 (0.62) | 0.52 (0.08) | 0.69 (0.07) | 0.75 (0.07) | 0.60 (0.06) |
| MIE No  | Nebout      | 0.50 (0.14) | -3.70 (2.43) | 1.97 (1.62) | 1.39 (0.56) | 0.47 (0.09) | 0.78 (0.07) | 0.80 (0.07) | 0.57 (0.10) |
|         | SAM-Resnet  | 0.29 (0.12) | -4.35 (2.73) | 2.39 (1.81) | 0.81 (0.40) | 0.37 (0.07) | 0.70 (0.09) | 0.73 (0.08) | 0.51 (0.08) |
|         | SAM-VGG     | 0.24 (0.13) | -4.82 (3.17) | 2.80 (2.05) | 0.70 (0.46) | 0.33 (0.06) | 0.62 (0.09) | 0.73 (0.08) | 0.52 (0.07) |
|         | Sa/GAN      | 0.29 (0.12) | -3.34 (0.90) | 1.83 (0.59) | 0.81 (0.41) | 0.36 (0.07) | 0.69 (0.09) | 0.74 (0.08) | 0.51 (0.09) |
|         | DeepGazeII  | 0.35 (0.12) | -4.25 (2.18) | 2.46 (1.68) | 0.96 (0.42) | 0.39 (0.09) | 0.66 (0.09) | 0.76 (0.06) | 0.51 (0.07) |
|         | MLNET       | 0.19 (0.12) | -2.45 (0.74) | 1.83 (0.52) | 0.50 (0.35) | 0.31 (0.08) | 0.62 (0.08) | 0.68 (0.07) | 0.51 (0.06) |

FIGURE 3. CC, NSS and AUC − B scores for the tested models (the x-axis). For each model, three scores are provided corresponding to the comparison with ICME dataset, MIE Fo dataset and MIE No dataset, respectively. This order aims to represent the degree of autism troubles, from high-functioning to Kanner syndrome.

FIGURE 4. The two-stream architecture of Nebout’s model [8].

we discuss the performances of these models over the three ASD eye tracking datasets presented in Section II.

1) PERFORMANCE OF NEUROTYPICAL MODELS

Table 4 presents the ability of neurotypical models to predict ASD saliency. Overall, results suggest that these models perform reasonably well for ICME dataset, decreases over the MIE Fo dataset and significantly drops down on MIE No dataset.

On ICME dataset, the best performances in terms of CC are obtained by DeepGaze II (CC = 0.73) and SAM-Resnet (CC = 0.72). The performance significantly drops down on MIE No dataset, where DeepGazeII and SAM-Resnet perform poorly (CC = 0.35, CC = 0.29, respectively). A similar significant trend is observed for all tested deep-based saliency models. Figure 3 illustrates graphically three similarity scores, i.e. CC, NSS and AUC − B computed over the three datasets. We observe that the CC scores drop down from 0.72 to 0.29, 0.60 to 0.24, 0.68 to 0.29, 0.73 to 0.35, 0.60 to 0.19 for SAM-Resnet, SAM-VGG, Sa/GAN, DeepGazeII and MLNET respectively.

This interesting observation may be explained by the degree of autism of participants. On the autism spectrum, participants involved in the ICME experiments present a low degree of autism (see Figure 2). This could explain the rather good ability of neurotypical models to predict ASD saliency maps. At the opposite of the spectrum, for participants of MIE Fo and MIE No who are mostly Kanner autisms, and more specifically for participants of MIE No, performances...
significantly drop down. This demonstrates that current neureotypical deep models are not appropriate and cannot generalize well to predict areas watched by people suffering from medium to high levels of disorders. In other words, current deep saliency models trained with eye-tracking data of neuretypical participants are only efficient to predict saliency for high-functioning autisms.

2) PERFORMANCE OF NEBOUT’s MODEL

To go further into the analysis, we evaluate the performance of Nebout’s model, a deep-based CNN model which has been trained with eye-tracking data of high-functioning autisms (ICME dataset). Results are given in Table 4.

Nebout’s model performs reasonably well in average for the three tested datasets. The model performs the best on ICME and MIE Fo datasets, with a correlation coefficient equal to 0.69 and 0.66, respectively. We also observe that the performance decreases with the degree of autism symptoms. The correlation coefficient decreases to 0.5 for MIE No. However, the performance drop is less pronounced than saliency models trained with neuretypical eye-tracking data. On MIE No, Nebout’s model significantly outperforms other models trained with ASD eye tracking data. On MIE No, Nebout’s model significantly outperforms other models trained with ASD eye-tracking data; the correlation coefficient, equal to 0.5, is much higher than the neuretypical models (the best performing neuretypical model performs in CC at 0.35). We also observe that the model performs better on MIE Fo than on MIE No data, and vice-versa. This result was not expected since Nebout model was trained with ASD eye-tracking data.

Table 5 gives the average rank for the tested models. The best possible rank is one. The higher the rank the less performing the model is. Nebout’s model ranks first for the three datasets. Figure 3 also shows that the performance of Nebout’s model is much higher than neuretypical model (see the third bar for each model) on MIE No dataset.

3) FINE-TUNING NEBOUT MODEL

Nebout’s model has been trained on MIT1003 dataset, and then fine-tuned with the eye-tracking data of ICME dataset. We propose to go one step further and to fine-tune this model with the MIE Fo eye tracking data and to test it on MIE No data, and vice-versa. The main idea is to evaluate the generalization degree of this model. If the model which has been fine-tuned on MIE Fo performs well on MIE No, it would suggest that the two sets of participants are very homogeneous. We proceed by using 20 images for fine-tuning the model and by testing it on 5 images; for each image, we have 17 and 12 scanpaths for MIE Fo and MIE No, respectively. As the number of training and testing images are very low, we do not aim to draw definitive conclusions. We are rather interested in the trend and variation of performance.

Table 6 presents the results. This shows that fine-tuning the model lowers its ability to predict ASD saliency. For instance, if we fine-tune the model with MIE No eye data, the performance of Nebout model on MIE Fo decreases. For instance, the CC score decreases from 0.66 to 0.54. A similar result is observed when the model is fine-tuned with MIE Fo. However, results also indicate that the difference is higher when the fine-tuning is performed with MIE No dataset.
the fine-tuning process (compared to MIE Fo). However, we believe that the degree of autism of MIE No is responsible for this performance drop. Indeed, there is a high discrepancy in visual deployment of the two ASD populations (MIE Fo and MIE No). Figure 6 illustrates differences between the human saliency maps computed from eye-tracking data stemming from MIE Fo and MIE No. Quantitatively speaking, the correlation coefficient $CC$ and the similarity $SIM$ scores computed by comparing the human saliency maps of MIE Fo and MIE No are equal to 0.53 and 0.49, respectively. These scores are rather low, supporting that the two sets of participants are very different.

**TABLE 7.** Performance of Nebout model trained with OSIE dataset, and with fine-tuning on ICME dataset.

| Dataset          | CC  | IG  | KL  | NSS | SIM | AUC-B | AUC-J | AUC-S |
|------------------|-----|-----|-----|-----|-----|-------|-------|-------|
| Without fine-tuning |     |     |     |     |     |       |       |       |
| ICME             | 0.56| -7.00| 5.08| 1.10| 0.57| 0.68  | 0.75  | 0.60  |
| MIE Fo           | 0.49| -7.68| 5.05| 1.09| 0.47| 0.68  | 0.74  | 0.57  |
| MIE No           | 0.33| -9.84| 6.14| 1.91| 0.27| 0.66  | 0.72  | 0.55  |
| With fine-tuning  |     |     |     |     |     |       |       |       |
| ICME             | 0.68| -1.48| 0.81| 1.26| 0.65| 0.77  | 0.79  | 0.61  |
| MIE Fo           | 0.63| -2.52| 1.35| 1.37| 0.58| 0.77  | 0.79  | 0.61  |
| MIE No           | 0.49| -3.18| 1.61| 1.34| 0.64| 0.76  | 0.80  | 0.57  |

fully consistent with our previous results presented in sections III-C1 and III-C2. The fine-tuning of Nebout model with ICME dataset significantly increases the performance over the three datasets spanning from low to high-functioning autism. This gain is especially important for the metrics $CC$, $IG$, $KL$, $NSS$ and $SIM$. For instance, the $CC$ scores significantly increase from 0.56 to 0.68 (paired t-test, $t(29) = -5.66, p < 0.05$), from 0.49 to 0.63 (paired t-test, $t(24) = -11.88, p < 0.05$) and from 0.33 to 0.49 (paired t-test, $t(24) = -15.37, p < 0.05$) for the ICME, MIE Fo and MIE No test datasets, respectively. For the different versions of $AUC$, the gain is more marginal. Although the fine-tuning process increases the performance, we are still quite far from performances of neurotypical saliency models on neurotypical eye-tracking data, as those shown in Table 3. Figure 7 illustrates some predicted saliency maps without and with fine-tuning for MIE Fo. The correlation coefficients are given in the caption and emphasizes the benefit to fine-tune with ASD data.

**D. TRAINING NEBOUT MODEL ON OSIE DATASET**

Because ICME dataset has some images in common with MIT300 dataset, a bias can be induced both during the training and the test. To make clear this point and to eliminate any ambiguity, we re-train Nebout model from a random initialization. For this purpose, we considered the neurotypical eye-tracking data of OSIE dataset [13]. We therefore split OSIE dataset in three sub-datasets: the training dataset consists of the first 400 images, which are augmented thanks to an horizontal flip. The next 100 images are used as the validation dataset, and the last 200 images as test dataset.

In order to predict where ASD people look at, the model is finally fine-tuned with data coming from the first 240 images of ICME dataset. The test dataset for ICME is then composed of 30 images. In addition, we use all images from MIE Fo and MIE No for the evaluation.

Table 7 presents the performance of two variants of Nebout model; one without the aforementioned fine-tuning and one with fine-tuning. This means we evaluate over the test datasets the performance of Nebout model when it has been trained with images from OSIE dataset but not finetuned (see wo finetuned). We then evaluate the model that has been finetuned with the training set of ICME dataset (see with finetuned).

From Table 7, several comments can be made. First we observe that, for both variants, the performance decreases with the severity of autism symptoms. The best performances are indeed obtained for the ICME dataset, involving high-functioning autism. The worst performances are again observed for the MIE No dataset. This is then

![FIGURE 6. Original images (top) and human saliency maps for MIE Fo (middle) and MIE No (bottom).](image)

**FIGURE 7.** Original images (top), predicted saliency maps with Nebout model trained with neurotypical data of OSIE dataset (middle) and predicted saliency maps with Nebout model trained with neurotypical data of OSIE dataset and fine-tuned with ICME data (bottom). From left to right, the $CC$ scores increase from 0.43 to 0.58, 0.55 to 0.64, 0.67 to 0.78 and 0.45 to 0.61.

**IV. DISCUSSION**

Saliency modelling consists in determining where observers look at. Most of existing approaches make strong assumptions, such as the existence of an unique saliency map in the brain, and that there exist a universal saliency map indicating where we look at. Such uniformity or universality of the saliency map presents some advantages since it provides a general overview of what attracts our attention. For many applications, for which the targeted or category of observers (e.g. age, gender, personal preference) is not known, it makes sense. However, we can go further.
Quite recently, some efforts have been made to change this paradigm and to put the observers in the midst of the design of saliency models. In other words, the observers become the key ingredient when it comes to simulate our visual behavior [31]–[33]. In [32], authors leverage personal preferences and image contents to tailor and to personalize the design of computational models of visual attention. In [33], authors leverage the differences in visual deployment between young children and adults in the prediction of saliency maps.

In this article, we focus on our ability to simulate the visual deployment of ASD people. Several attempts have been proposed recently such as [7]–[9], [12]. We however question the appropriateness of modelling the visual behavior of ASD people. Our first answer is that it really makes sense to go further in this direction and to pursue our effort in the modelling. This aims to provide a better understanding of how attention is deployed in people with ASD, especially when natural scenes are used; natural stimuli provide a greater ecological validity. In addition, the recent deep-based saliency networks and their activation maps [34] would allow us to characterize the various visual attributes that could influence saliency [2], [11].

However, we need to be careful on our claims. According to our results, an universal saliency model indicating where ASD people look at does not seem reasonable due to the variety and the diversity of autism troubles. The autism spectrum is so large and different, that it is required to personalize as much as possible our future design of saliency model.

We are witnessing a growing interest in the computational modelling of visual attention of ASD people. The most recent methods relying on eye-movement recording allow us to collect a large amount of data, which could be used to train neural networks. Visualizing the visual features learned by the network may help us to understand the factors influencing the visual attention. However, as the differences within the population of ASD people are so large, it is extremely difficult to draw definitive conclusions from a given study. This article aims to point out the huge variety of autism and then the difficulty to determine the factors influencing the visual attention of ASD people. We, i.e. computer scientists working at the frontier of computer science and psychiatry, have to understand and adapt the computational modelling to the degree of autism of subjects.

In future works, the community has to contribute for increasing the coverage of the autism spectrum in terms of eye-tracking data. This means that Kanner’s and Asperger’s syndromes which are at the two ends of the same spectrum must be considered to account for the large variety of autism. For Kanner autism, the severity of the symptoms is very high. Symptoms consist of a profound lack of affect or emotional contact with others, an intense wish for sameness in routines, muteness or abnormality of speech, a fascination with manipulating objects as described in [18]. Performing eye-tracking experiments with such population is extremely difficult and would require to define and to standardize an ecological eye-tracking protocol in order to ensure both the well-being of participants and the success of the experiment. New data processing procedures might also be required because of the poor visual engagement of Kanner people. By increasing the number of participants and their variety going from Asperger to Kanner syndrome, we will be able to understand the visual features that influences the visual deployment of ASD people along the autistic spectrum.

REFERENCES

[1] J. Baio et al., “Prevalence of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2014,” MMWR Surveill. Summaries, vol. 67, no. 6, p. 1, 2018.
[2] C. Ames and S. Fletcher-Watson, “A review of methods in the study of attention in autism,” Develop. Rev., vol. 30, no. 1, pp. 52–73, Mar. 2010.
[3] L. M. Schmitt, E. H. Cook, J. A. Sweeney, and M. W. Mosconi, “Saccadic eye movement abnormalities in autism spectrum disorder indicate dysfunction in cerebellum,” Brain Res. Bull., vol. 5, no. 1, p. 47, 2014.
[4] E. Loth, “Abnormalities in ‘cultural knowledge’ in autism spectrum disorders. A link between behavior and cognition?” in Autism: An Integrated View From Neurocognitive, Clinical, and Intervention Research. Hoboken, NJ, USA: Wiley, 2007.
[5] J. Bradshaw, F. Shic, A. N. Holden, E. J. Horowitz, A. C. Barrett, T. C. German, and T. W. Vernon, “The use of eye tracking as a biomarker of treatment outcome in a pilot randomized clinical trial for young children with autism,” Autism Res., vol. 12, no. 5, pp. 779–793, May 2019.
[6] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. L. Callet, “A dataset of eye movements for the children with autism spectrum disorder,” in Proc. 10th ACM Multimedia Syst. Conf., Jun. 2019, pp. 255–260.
[7] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai, “Visual attention analysis and prediction on human faces for children with autism spectrum disorder,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 15, no. 3s, pp. 1–23, Jan. 2020.
[8] A. Nebout, W. Wei, Z. Liu, L. Huang, and O. Le Meur, “Predicting saliency maps for ASD people,” in Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW), Jul. 2019, pp. 629–632.
[9] W. Wei, Z. Liu, L. Huang, A. Nebout, and O. Le Meur, “Saliency prediction via multi-level features and deep supervision for children with autism spectrum disorder,” in Proc. IEEE Conf. Multimedia Expo Workshops (ICMEW), Jul. 2019, pp. 621–624.
[10] M. Jiang and Q. Zhao, “Learning visual attention to identify people with autism spectrum disorder,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 3267–3276.
[11] S. Wang, M. Jiang, X. M. Duchesne, E. A. Laugues, D. P. Kennedy, R. Adolphs, and Q. Zhao, “Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking,” Neuron, vol. 88, no. 3, pp. 604–616, Nov. 2015.
[12] L. Q. Uddin, K. Supekar, C. J. Lynch, A. Khouzam, J. Phillips, C. Feinstein, S. Ryali, and V. Menon, “Saliency network–based classification and prediction of symptom severity in children with autism,” JAMA Psychiatry, vol. 70, no. 8, pp. 869–879, 2013.
[13] J. Xu, M. Jiang, S. Wang, M. S. Kankanahalli, and Q. Zhao, “Predicting human gaze beyond pixels,” J. Vis., vol. 14, no. 1, pp. 1–20, 2014.
[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look in,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep. 2009, pp. 2106–2113.
[15] B. Follet, O. Le Meur, and T. Baccino, “New insights into ambient and focal visual fixations using an automatic classification algorithm,” i-Perception, vol. 2, no. 6, pp. 592–610, 2011.
[16] O. Le Meur and P.-A. Fons, “Predicting image influence on visual saliency distribution: The focal and ambient dichotomy,” in Proc. ACM Symp. Eye Tracking Res. Appl., Jun. 2020, pp. 1–5.
[17] American Psychiatric Association. (2013). Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Pub.
[18] L. Kanner, “Autistic disturbances of affective contact,” Nervous Child, vol. 2, pp. 217–250, Apr. 1943.
[19] D. D. Salvucci and J. H. Goldberg, “Identifying fixations and saccades in eye-tracking protocols,” in Proc. Symp. Eye Tracking Res. Appl. (ETRA), 2000, pp. 71–78.
OLIVIER LE MEUR (Member, IEEE) received the Ph.D. degree from the University of Nantes, in 2005. From 1999 to 2009, he has worked in the media and broadcasting industry. In 2003, he joined the Research Center of Thomson-Technicolor, Rennes, where he supervised a research project concerning the modeling of the human visual attention. Since 2009, he has been supervising the research team PERCEPT/IRISA. His research interest includes understanding of the human visual attention, including computational modelling of the visual attention and saliency-based applications (video compression, objective assessment of video quality (SDR, HDR), and retargeting).

ALEXIS NEBOUT received the master’s degree in computer sciences from the École Supérieure d’Ingénieurs de Rennes (ESIR), in 2019. He contributes to the design of saliency models for determining where ASD people look at. His main research interest includes computational modeling of visual attention.

MYRIAM CHEREL received the Ph.D. degree from the University of Rennes, in 2007. She has worked in MIEs with autistic children for 12 years. She became an Associate Professor [Section CNU 16 (Psychopathologie clinique et clinique psychanalytique)] with the University of Rennes 2, in 2008. She is in charge of the master 2 of « Psychopathologie de l’adulte ». She has been supervising the Research Group on Autism (GRA), since 2010. She is a member of the research laboratory Recherches en Psychopathologie et Psychanalyse (RPpsy). She is the author of Affinity therapy Nouvelles recherches sur l’autisme (PUR, 2015) and Etre parents au 21ème siècle: Des parents rencontrent des psychanalystes (Editions Michèle, 2018).

ELISE ETHCHAMENDY received the master’s degree in psychology (option clinical and psychoanalytical psychopathology) from the University of Rennes, in 2020. During her internship, she has investigated the affinity therapy on people suffering from ASD. She focused more specifically on the visual engagement of ASD people while viewing neutral or affinity images.

***