

Attention-based Autism Spectrum Disorder Screening with Privileged Modality

Shi Chen

Qi Zhao

Department of Computer Science and Engineering,
University of Minnesota

chen4595@umn.edu

qzhao@cs.umn.edu

Abstract

This paper presents a novel framework for automatic and quantitative screening of autism spectrum disorder (ASD). It is motivated to address two issues in the current clinical settings: 1) short of clinical resources with the prevalence of ASD (1.7% in the United States), and 2) subjectivity of ASD screening. This work differentiates itself with three unique features: first, it proposes an ASD screening with privileged modality framework that integrates information from two behavioral modalities during training and improves the performance on each single modality at testing. The proposed framework does not require overlap in subjects between the modalities. Second, it develops the first computational model to classify people with ASD using a photo-taking task where subjects freely explore their environment in a more ecological setting. Photo-taking reveals attentional preference of subjects, differentiating people with ASD from healthy people, and is also easy to implement in real-world clinical settings without requiring advanced diagnostic instruments. Third, this study for the first time takes advantage of the temporal information in eye movements while viewing images, encoding more detailed behavioral differences between ASD people and healthy controls. Experiments show that our ASD screening models can achieve superior performance, outperforming the previous state-of-the-art methods by a considerable margin. Moreover, our framework using diverse modalities demonstrates performance improvement on both the photo-taking and image-viewing tasks, providing a general paradigm that takes in multiple sources of behavioral data for a more accurate ASD screening. The framework is also applicable to various scenarios where one-to-one pairwise relationship is difficult to obtain across different modalities.

1. Introduction

Autism spectrum disorder (ASD) is a heritable and life-long neurodevelopmental disorder (NDD) with complicated

aetiology and causes. It is globally prevalent, and affects one in 59 children in the United States [2]. Though currently recognized as the most effective clinical route to ASD treatment [3], early diagnoses and interventions rely on a team of medical expertise with diagnostic instruments that are both time-consuming and clinically demanding. Due to the prevalence of ASD and limited clinical resource, they are not widely applicable. In addition, human assessment is subjective and tends to be inconsistent, and is also episodic. As a result, automatic and objective tools that assist ASD screening have been of significant clinical and societal need.

The visual attention network is pervasive in the brain that many NDDs are associated with atypical attention towards visual stimuli. For example, people with ASD have been long known to have atypical attention to faces or other social stimuli [6, 7, 27, 28, 29]. Recent studies with natural scene stimuli show more complicated or finer differences between people with ASD and healthy people [36]. This paper **develops novel computer vision techniques to address existing challenges in ASD screening**. It proposes a new approach that allows to record and model attentional preference with greater ecologically validity and practical feasibility. In addition, with the complexity of the problem, *e.g.* considerable heterogeneity within ASD or across multiple NDDs [24]), and the scarcity of clinical data, it highlights the importance and proposes methods to make use of multiple behavioral modalities as well as temporal information to encode more detailed and comprehensive information needed for accurate ASD screening.

Specifically, we propose to incorporate two distinct modalities related to human visual attention, *i.e.* attentional preference recorded from a photo-taking task and an image-viewing task, for ASD screening. In the photo-taking task, subjects freely move in the environment and identify their preferred regions of interest by taking photos, while in the image-viewing task, subjects view different images with their eye movements recorded by an eye-tracking device. Instead of screening ASDs independently on each modality, we present a novel ASD screening with privileged modality framework that integrates diverse modalities during training

and benefits each modality at testing. Our framework consists of three principal components: to leverage the pervasiveness of visual attention network and the learning potential of deep neural networks (DNNs), we develop two DNN models each encoding the characteristics of photos taken by subjects (photo-taking) or the temporal information of eye movements (image-viewing) to classify ASD based on attentional preference. To make use of the abundant and complementary information from the two modalities, we propose a multi-modal distillation method that learns a shared embedding space for multiple modalities (*i.e.* main modality available all the time and privileged modality applicable only during training) and distills multi-modal knowledge from the shared space to each modality. Compared to existing methods, our framework is advantageous in: 1) unlike previous methods [17, 22] that pay less attention to the temporal information and independently train the feature encoder and the ASD classifier, our image-viewing model is developed in an end-to-end manner and takes in the temporal information of eye movements; 2) different from the multi-modal methods [21, 34] that require the availability of all modalities during testing, which is difficult to be fulfilled in clinical scenarios, our framework can be deployed on each independent modality; 3) instead of relying on the one-to-one pairwise relationship between modalities similar to the other learning with privileged modality methods [10, 16, 20, 23], the proposed multi-modal distillation method can transfer abundant information across different modalities without overlap in subjects.

In summary, this paper carries three major contributions:

- We go beyond one modality and present an ASD screening with privileged modality framework that utilizes multiple source of behavioral data. In our context, there is no subject overlap between modalities.
- We develop the first computational model to screen ASD based on a photo-taking task. Despite the challenging nature of the task, our model outperforms human experts and achieves reasonable performance.
- By incorporating temporal information of eye movements, our model on image-viewing task is able to achieve the new state-of-the-art performance.

2. Related Works

Automatic ASD screening. There are several computational models that automatically identify people with ASD. Anzulewicz *et al.* [1] use smart tablet devices to record the motor patterns of children, and propose three decision-tree based models for identifying ASD based on these patterns. Inspired by the findings that individuals with ASD have difficulty recognizing faces and interpreting facial emotions [29], Liu *et al.* [22] evaluate the face scanning patterns

of children and detect those with ASD. To capture differences of gaze patterns between ASD and control group during image-viewing, Wang *et al.* [36] propose to train a support vector machine (SVM) model with pre-defined features to classify individuals with ASD based on their gaze patterns. Jiang and Zhao [17] later on extend the ideas of [36] by introducing a new deep neural network approach that highlights the differences of gaze patterns, resulting in more discriminative features for accurate ASD screening. People have also explored different types of neuroimaging techniques for classifying ASD [21, 34]. While these methods achieve reasonable results, they either consider only a single modality and pay less attention on utilizing temporal information [1, 17, 22, 36], or rely on multi-modal data acquired by resource-demanding instruments that are difficult to deploy in clinical scenarios [21, 34].

Learning Under Privileged Information. Learning under privileged information (LUPI) is a paradigm proposed in [31, 32], specifying the scenarios where certain privileged information is available during training but inapplicable at testing. In this paper, we focus on the case that the privileged information corresponds to a modality different from the one available all the time, *i.e.* learning with privileged modality. Hoffman *et al.* [16] propose a multi-stream hallucination architecture that learns the mappings between different modalities and emulates the multi-modal scenario at testing with a single modality. In [20], Lambert *et al.* make use of features from a privileged modality to learn the hyper-parameters for dropout units. Garcia *et al.* [10] propose to incorporate modality hallucination [16] with knowledge distillation [14] for action recognition with privileged modality. In [23], Luo *et al.* distill multi-modal knowledge to a single-modal network by constructing a graph structure between various modalities during training. These methods rely on one-to-one pairwise relationship between different modalities, which is typically difficult to fulfill in the clinical settings with patient data, *e.g.* in our ASD screening experiments the data for two modalities are collected separately from two groups of subjects without overlap.

In this work, we propose an ASD screening with privileged modality framework that incorporates multiple behavioral modalities for ASD screening. Our framework does not rely on subject overlap across different modalities or advanced diagnostic instruments for data collection, thus is suitable for ASD screening in regular clinical settings. By incorporating temporal information of eye movements and abundant knowledge from the two modalities, the proposed models are able to achieve superior performance and outperform the previous state-of-the-art methods.

3. ASD Screening with Privileged Modality

Neurodevelopmental disorders, such as ASD, are typically characterized by multiple symptoms, where a single

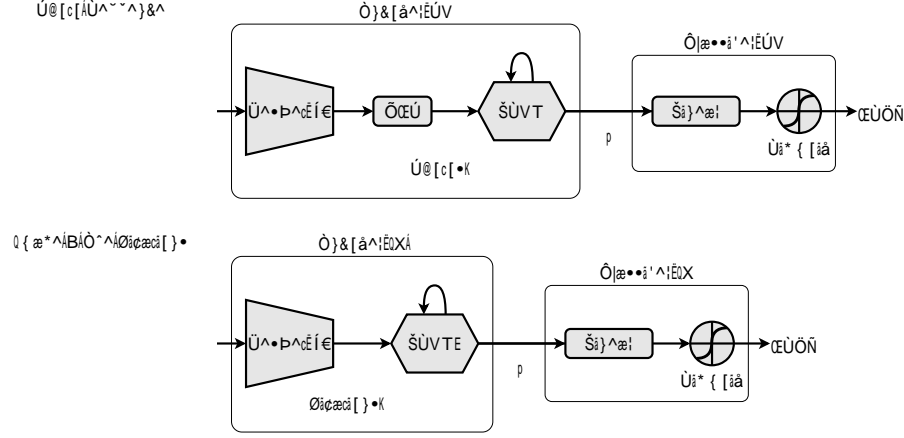


Figure 1: High-level architectures for attention based ASD screening models on photo-taking (top) and image-viewing (bottom) modalities. GAP denotes the global average pooling layer. \hat{x}_t in photo-taking is the features for image t , while in image-viewing \hat{x}_t is the features extracted at the proximity of fixation t . N and M represent the number of images and fixations in photo-taking and image-viewing data.

modality may not carry sufficient information for diagnostic purposes. In this work, we propose to screen ASD with more behavioral modalities that will provide complementary and abundant information. Multi-modality data is especially important with the heterogeneity of the conditions and the scarcity of data from clinical populations.

Specifically, we present an ASD screening with privileged modality framework that utilizes two distinct behavioral modalities, *i.e.* photo-taking and image-viewing. It incorporates information from both modalities during training and only requires one modality at testing, *i.e.* we treat one modality as the privileged modality that benefits the learning of another main modality. Unlike existing multi-modal [21, 34] or learning with privileged modality [10, 16, 20, 23] methods, our framework does not rely on the availability of all modalities during deployment or one-to-one pairwise relationship (*e.g.* subject overlap) across modalities, making it more practical for real-world clinical scenarios.

In this section, we illustrate the three major components of the proposed framework, including two DNN models for ASD screening on photo-taking and image-viewing task respectively, and a multi-modal distillation method that distills multi-modal knowledge from a shared space to each independent modality.

3.1. ASD Screening on Photo-Taking

Different from previous visual attention based ASD screening methods [17, 22, 36] that follow a passive image-viewing procedure, our photo-taking task allows subjects to freely interact with various scenarios and identify regions or objects of interest in the first-person settings, providing a more ecological paradigm in revealing one’s attentional preference. Moreover, photos taken by people offer addi-

tional information that displays their behaviors on social interactions. For example, due to reduced social interactions, individuals with ASD may not ask people for pose adjustment, resulting in poor-quality photos with people not posing or looking at the camera. Inspired by the findings that photo taken by people with ASD tend to have different characteristics from those taken by healthy people [35], *e.g.* difference in attentional preference and in quality of photos, in this paper we aim at screening ASD via characterizing these differences with a pool of photos taken by the subjects.

To accomplish the aforementioned objective, we propose to leverage a CNN for learning meaningful features and a Recurrent Neural Network (RNN) for capturing the characteristics of a sequence of photos. As shown in Figure 1 (top), the proposed model consists of two major components: 1) an **encoder module** that first projects raw image data to **high-level visual features using the state-of-the-art ResNet-50** [13], and then sequentially traverses the features for different images within a photo sequence using a **Long Short Term Memory (LSTM)** [15] network, and 2) a classifier module that takes in the final hidden state of LSTM and makes the prediction (*i.e.* ASD or Control). Given the photo sequence taken by a specific subject, we first compute the visual features for different images within the sequence via ResNet-50, and then apply global average pooling (GAP) to convert the spatial features to vectors that describe the abstract information about the corresponding images. These vectors are then sequentially forwarded to the LSTM, capturing the characteristics of photos by repeatedly updating the hidden state. After obtaining the final hidden state which encodes information of the whole sequence, we directly feed it to our classifier (*i.e.* a single fully-connected layer) for identifying people with ASD.

Figure 2: Comparison of the gaze patterns between ASDs and Controls. From left to right are fixation maps for four continuous time steps, and the aggregated fixation maps.

3.2. ASD Screening on Image-viewing

Recent advances in ASD research using naturalistic scenes have led to several new insights in how people with ASD look differently with healthy subjects. For example, with complex stimuli containing rich social and semantic content, Wang *et al.* [36] observe that the order of fixations and latency to semantic objects differ significantly between subject groups, suggesting the role of temporal information in the screening task. While previous works [17, 22] have investigated the feasibility of classifying ASD with visual attention, limited effort is placed on exploring the effectiveness of temporal information encoded within the eye movements. Figure 2 highlights the importance of using temporal information which reveals significant difference of gaze patterns between ASDs and controls, even though the aggregated fixation maps are similar. Moreover, due to the scarcity of clinical data which prevents over-complicated model designs, these methods typically train a feature encoder and an ASD classifier separately without explicitly correlating the learned visual features with ASD screening, making it difficult for them to achieve satisfying performance. We in this section introduce a DNN model to address these issues for more accurate ASD screening with image-viewing. Our model is optimized in an end-to-end fashion, which automatically connects visual features with ASD screening, and takes in temporal information of eye movements to decipher more discriminative features of visual attention recorded during image-viewing.

As depicted in Figure 1 (bottom), the proposed image-viewing model shares a similar design as our model for the photo-taking modality. However, unlike photo-taking where the goal is to classify ASD based on a sequence of photos, for image-viewing we aim at differentiating people with ASD based on their attention pattern (*i.e.* eye movements) captured on each specific image. As a result, given an image and its corresponding visual scanpath for a specific subject, instead of utilizing a CNN together with GAP

for feature extraction, we first obtain useful visual features from the CNN and then extract features at the proximity of each eye fixation (*i.e.* 2048-dimensional feature vector at the closest location to each fixation). The extracted features are then sequentially fed to our LSTM (note that for image-viewing we use a variant of LSTM similar as [11] for better performance, denoted as LSTM) based on the order of fixations within the scanpath, capturing the temporal information of the eye movements. At each eye fixation, the process can be represented as follows:

$$\mathbf{i}_t = (\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{W}_{ic}\mathbf{c}_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = (\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{W}_{fc}\mathbf{c}_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{o}_t = (\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{W}_{oc}\mathbf{c}_{t-1} + \mathbf{b}_o) \quad (3)$$

$$\mathbf{m}_t = \tanh(\mathbf{W}_{mx}\mathbf{x}_t + \mathbf{W}_{mh}\mathbf{h}_{t-1} + \mathbf{b}_m) \quad (4)$$

$$\mathbf{c}_t = \mathbf{i}_t \odot \mathbf{m}_t + \mathbf{f}_t \odot \mathbf{c}_{t-1} \quad (5)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \mathbf{c}_t \quad (6)$$

where \mathbf{x}_t is the visual features extracted at the proximity of the t_{th} eye fixation, \mathbf{W} and \mathbf{b} the trainable parameters in LSTM, the sigmoid function, and \mathbf{h}_{t-1} and \mathbf{c}_{t-1} represent the hidden state and memory vector that contains temporal information of previous eye movements. \mathbf{i} , \mathbf{f} and \mathbf{o} are input gate, forget gate and output gate for LSTM, and \mathbf{m} further encodes features based on \mathbf{x}_t and \mathbf{h}_{t-1} . The hidden state \mathbf{h} computed at the end of the visual scanpath is fed into the classifier for predicting people with ASD.

3.3. Multi-Modal Distillation through Shared Space

With the aforementioned models for ASD screening on each modality, a key here is to effectively integrate the information from the two distinct modalities to further improve the performance of ASD screening on each of them. For that, we propose a multi-modal distillation method that enables models to learn from diverse types of behavioral data through a shared space. Our method is inspired by the cross-modal retrieval and matching methods [5, 12, 25], however, it significantly differs from them in both goal and methodology: 1) different from [5, 12, 25] whose goal is to retrieve samples in the target modality with data in the source modality, our aim is to create a shared space for transferring knowledge across different modalities for performance improvement on each modality, and 2) in the cross-modal matching, *e.g.* [5], modules of different modalities are optimized under a multi-task learning framework [4] (joint training on different modalities), where achieving satisfying performance on both modalities is difficult. Instead, we propose a novel method that distills multi-modal knowledge through jointly training the shared space, but overcomes the mentioned difficulty by disentangling models of different modalities after learning the shared space, so each model could focus on its own modality to best optimize it. Figure 3 shows the procedure of the proposed method.

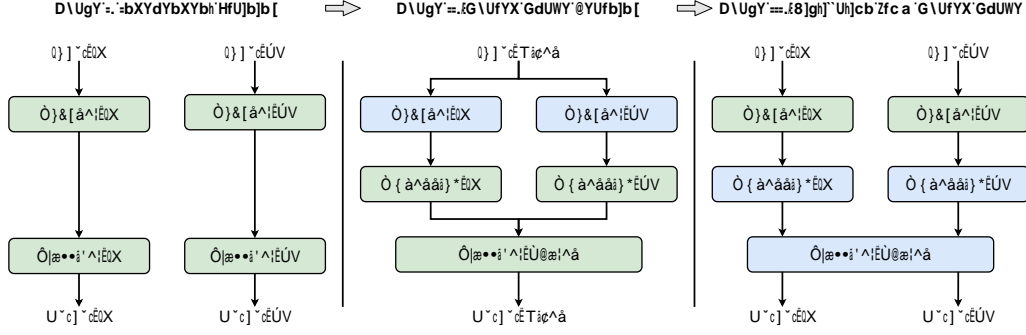


Figure 3: Process for the proposed ASD screening with privileged modality framework. Different training phases are highlighted with bold text at the top. Encoder-IV, Encoder-PT, Classifier-IV and Classifier-PT are the same as those presented in Figure 1. Modules with blue color are fixed during a training phase while those colored in green are being optimized.

Our method follows the intuition of first constructing a shared space that encodes multi-modal knowledge and then encouraging modules of each modality to learn from such space. Specifically, to develop a shared embedding space with sufficient understanding on each of the modalities, our method first independently optimizes the models on their corresponding modalities (Independent Training). With the learned modal-specific knowledge, we then integrate models of each modality (Shared Space Learning) and construct the shared space by jointly training on the two modalities with the following loss function L :

$$L = \text{BCE}(Y_I, \hat{Y}_I) + \text{BCE}(Y_P, \hat{Y}_P) \quad (7)$$

$$[\hat{Y}_I, \hat{Y}_P] = W_{\text{cls}}[W_I X_I, W_P X_P] \quad (8)$$

where Y_I and Y_P are ground truth annotations of the image-viewing and photo-taking modalities, \hat{Y}_I and \hat{Y}_P are the respective model predictions, and BCE represents binary cross-entropy loss. X_I and X_P are features extracted by the encoders of two modalities (*i.e.* Encoder-IV and Encoder-PT in Figure 3), W_I and W_P denote the embedding layers for the two modalities (*i.e.* Embedding-IV, Embedding-PT), and W_{cls} corresponds to the shared classifier (*i.e.* Classifier-Shared). By fixing the modal-specific modules (*i.e.* Encoder-IV and Encoder-PT) and only optimizing the embedding layers as well as the shared classifier (*i.e.* Embedding-IV, Embedding-PT and Classifier-Shared) using the above equations, we construct the shared space by learning essential knowledge from both modalities.

In order to distill multi-modal knowledge from the shared space to each modality while alleviating the training difficulties in multi-task learning, instead of continuing the joint training [4, 5], we propose to disentangle the models of different modalities and optimize them separately on their own modalities. Specifically, during the Distillation from Shared Space phase, the embedding layers as well as the shared classifier (*i.e.* Embedding-IV, Embedding-PT and Classifier-Shared) are fixed and only modules of each in-

dependent modality (*i.e.* Encoder-IV and Encoder-PT) are optimized, encouraging the modules of each modalities to adapt to the shared space with multi-modal knowledge and learn aligned feature representation on the two modalities.

The aforementioned procedure connects different modalities via learning a shared space, and encourages models to distill multi-modal knowledge from it to improve the performance of each modality. In our context, with the same classifier shared among both modalities, our method learns aligned feature representation across the two behavioral modalities for ASD screening, allowing them to complement each other and mutually boost their feature representations with the multi-modal knowledge encoded in the shared space. We note that the proposed method is applicable to scenarios where only partial modalities (in our case only one modality) are available at testing and one-to-one pairwise relationship across modalities does not exist, common in the clinical settings. Section 4.2 demonstrates that the proposed method with privileged modality can improve the performance of ASD screening models on two distinct modalities, while Section 4.3 analyzes the knowledge learned in the shared space.

4. Experiments

In this section, we report implementation details and a comprehensive evaluation of the proposed methods.

4.1. Implementation

Datasets and Evaluation. We first introduce data used in this work. For photo-taking, 22 individuals with ASD and 23 controls (*i.e.*, healthy people with matched age, gender and IQ) participated in our experiments. They were instructed to take photos in both indoor and outdoor scenarios, and each took 40 photos on average. For image-viewing, we use eye-tracking data from 20 ASDs and 19 controls. Binocular eye movements were recorded while viewing 700 images from the OSIE [37] eye-tracking dataset, and we

treat different eyes of a patient as two subjects in the evaluation similar as [17]. Eye fixations were extracted using Cluster Fix method [19]. Note that there is no subject overlap between two modalities. To show the generality of our methods, we also conducted experiments on the recent Saliency4ASD [9] dataset.

For evaluation, we adopt the widely used leave-one-subject-out cross-validation as in [17, 22], which is capable of returning an almost unbiased estimate of the probability of error [33]. Note that since Saliency4ASD [9] does not provide subject IDs, we use the i_{th} fixation sequences from the same group (*i.e.* ASD or control) in the training data to construct the validation data for the i_{th} round of cross validation. Similar as [17, 22], we evaluate our models with accuracy, sensitivity (*i.e.* true positive rate), specificity (*i.e.*, true negative rate) and Area Under the ROC Curve (AUC).

Model Specification. ResNet-50 [13] used in our models are first pre-trained on ImageNet dataset [8] and then jointly optimized with other modules in both models. The embedding size for LSTMs (LSTM* and LSTM) is 512, while for Embedding-IV and Embedding-PT (see Figure 3) in the multi-modal distillation we set their size to 1024. For image-viewing, we use the original image together with all of its corresponding eye fixations as input, while for photo-taking we randomly sample 12 photos (number of photos set based on empirical results) from the photo pool of a specific subject as a single input sequence.

Training. In order to train our attention based ASD screening models, for photo-taking we traverse samples for all subjects (excluding the one for validation) and each subject is associated with a photo sequence randomly sampled from his photo pool. While for image-viewing, we use the same image selection technique from [17] to select the top-100 discriminative images that best differentiate gaze patterns between ASDs and Controls with training subjects for each round of validation. We utilize Adam [18] optimizer with binary cross-entropy loss to train all of our models using weight decay 10^{-5} and gradient clipping 10. The batch sizes for both tasks are set to 12. During independent training, the models are trained for 10 and 180 epochs for image-viewing and photo-taking, with learning rate initialized as 10^{-4} and divided by 2 every 2 and 30 epochs. To learn the shared space, we jointly optimize the models on both modalities with learning rate 5×10^{-6} and a single epoch (since the datasets for two modalities have different sizes, we continuously train the models until data for both modalities are processed). After successfully learning the shared space, we separately train the models on image-viewing and photo-taking (the Distillation from Shared Space phase in Figure 3) for 3 and 60 epochs respectively.

Subject-wise Classification. Since the evaluations of ASD screening are performed on a subject basis, to convert our sample-wise predictions (prediction on different images

	Acc.	Sen.	Spe.	AUC
Liu <i>et al.</i> [22]	0.89	0.93	0.86	0.89
Jiang <i>et al.</i> [17]	0.92	0.93	0.92	0.92
IV-Independent	0.97	1.00	0.95	1.00
IV-Full	0.99	1.00	0.98	1.00
IV-Independent (Saliency4ASD)	0.89	0.86	0.93	0.92
IV-Full (Saliency4ASD)	0.93	0.93	0.93	0.98
Human Expert [35]	0.65	-	-	-
PT-Independent	0.76	0.77	0.74	0.82
PT-Full	0.84	0.77	0.91	0.84

Table 1: Inter-model comparison on ASD screening. Results on our image-viewing dataset, Saliency4ASD [9] and our photo-taking dataset are divided by the horizontal lines and listed from top to bottom. IV-Independent and PT-Independent are our single-modal models on image-viewing and photo-taking. Our full models with multi-modal distillation are denoted as IV-Full and PT-Full for both modalities. Four evaluation metrics are used, including Accuracy (ACC.), Sensitivity (Sen.), Specificity (Spe.) and AUC. Best results are highlighted in bold text.

or photo sequences) to subject-wise predictions, we average the confidences of all samples for a subject (top-100 discriminative images for image-viewing and 5 randomly sampled sequences for photo-taking) and utilize a pre-defined threshold, *i.e.* 0.5, to identify ASD.

4.2. Results

In this section, we report the experimental results to demonstrate the effectiveness of our ASD screening with privileged modality framework. We first perform inter-model comparison between the proposed models and the related state-of-the-art. Specifically, for image-viewing we compare our model with [17, 22] which also screen ASD based on eye fixations, while for photo-taking we use human expert performance [35] (percentage of three human experts agreeing on the same labels, *i.e.* ASD or control) as a reference as this is the first computational model on this task. We then conduct intra-model comparison on our models at different phases of the proposed multi-modal distillation method. Table 1 and Table 2 show the quantitative results on inter-model and intra-model comparisons.

As shown in Table 1, the proposed model on the image-viewing modality with temporal information is able to significantly outperform the current state-of-the-art ASD screening models by all evaluation metrics. By using a recurrent module to scan through visual features at different eye fixations in their temporal order, our model achieves 100% accuracy in recognizing individuals with ASD (sensitivity) and 95% accuracy on discriminating healthy people (specificity). Identifying people with ASD on photo-taking is more challenging and the performance of human experts is 65%. Our model shows reasonable performance when training with only its own modality (76% overall accuracy).

	Acc.	Sen.	Spe.	AUC
IV-Independent	0.97	1.00	0.95	1.00
IV-Shared	0.97	1.00	0.95	1.00
IV-Full	0.99	1.00	0.98	1.00
IV-Extra	0.97	1.00	0.95	1.00
PT-Independent	0.76	0.77	0.74	0.82
PT-Shared	0.78	0.77	0.78	0.82
PT-Full	0.84	0.77	0.91	0.84
PT-Extra	0.73	0.73	0.74	0.82

Table 2: Intra-model comparison of the proposed multi-modal distillation. Within each section of a specific modality, the first three results correspond to models after different training phases (see Figure 3, results are arranged in the same order as training phases), while the last row, *i.e.* -Extra, shows single-modal performance with additional layers (Embedding-IV or Embedding-PT in Figure 3) and the same amount of training epochs as -Full.

Figure 4: t-SNE visualization of features extracted at the three different training phases of the proposed multi-modal distillation method on the photo-taking modality. From (a) to (c) is the result on Independent Training, Shared Space Learning and Distillation from Shared Space phase. Red dots represent samples for ASD, while blue dots correspond to samples for Control.

Moreover, by incorporating the two modalities with the proposed multi-modal distillation method, we are able to achieve considerable improvements on both modalities. Specifically, we further boost the overall accuracy from 97% (single-modal performance) to 99% for image-viewing, and significantly increase the overall accuracy from 76% to 84% for photo-taking.

According to the intra-model comparisons reported in Table 2, the performance of ASD screening is increased monotonically across the three phases. Particularly, with the shared space constructed in the Shared Space Learning phase, our multi-modal distillation method significantly improves the performance by distilling multi-modal knowledge to modules of each independent modality in the Distillation from Shared Space phase. We further look into the aligned features learned from the multi-modal knowledge, and compare the features learned at the three phases (photo-taking modality) using t-SNE [30] visualization. As shown in Figure 4, features learned solely from a single modality are not sufficiently discriminative, thus data points with dif-

ferent labels (ASD or Control) are mixed together. After the Shared Space Learning, samples for different labels begin to move towards separate directions. Finally, by transferring multi-modal knowledge from the shared space to modules of independent modality, the aligned features become more discriminative and are well separated into two clusters.

To validate the contributions of the proposed multi-modal distillation method, we train our models on each single modality but with additional layers (Embedding-IV or Embedding-PT in Figure 3) and the same amount of training epochs as multi-modal distillation. We denote this method as -Extra in Table 2. Results show that additional layers and training on a single modality has non-significant (image-viewing) or even negative (photo-taking) effects. The results confirm that distilling knowledge across different modalities using the proposed method plays an essential role in boosting the performance of attention based ASD screening, and our improvements are not merely due to the advantages of model modifications or extra training epochs.

4.3. What Did the Shared Space Learn?

So far we have demonstrated that, by learning a space shared across the two modalities and distilling multi-modal knowledge from the space to modules of each modality, we are able to improve the accuracy of ASD screening by a considerable margin. To shed more light on the effectiveness of our multi-modal distillation method, in this section we focus on analyzing the knowledge learned in the shared space through both qualitative and quantitative evaluations. More specifically, we study what the shared space learn on correlating the two modalities and why it is able to benefit ASD screening on each independent modality.

Qualitative Evaluation. To understand how the shared space correlates the two modalities, we first extract the features computed by modules of different modalities in the independent space (features computed in the Encoder-IV and Encoder-PT after the Independent Training phase) and that in the shared space (features computed at Embedding-IV and Embedding-PT after the Shared Space Learning phase), and then match the nearest inputs between the two modalities based on their corresponding features. We use cosine similarity as the distance metric for matching the nearest inputs, which is widely used in Natural Language Processing for matching different meaningful words [26]. By comparing the nearest inputs between the independent and shared space, we are able to reveal the alignments of modalities in the shared space. Figure 5 shows qualitative results of the matched examples, *i.e.* nearest inputs between photo-taking (entire photo sequences, each has 12 photos taken by one subject) and image-viewing (images with fixated regions being highlighted). Each row represents one pair of match. Note that in image-viewing, only the fixated regions are compared. We make three key observations as follows:

Figure 5: Nearest inputs between photo-taking (photo sequences, left) and image-viewing (images with fixated regions, right), best viewed in digital form with zooming. For each row, from left to right are entire photo sequences (each has 12 images in total, displayed on 1st-12th columns), matched counterparts from the image-viewing modality in the independent (13th column) and the shared space (14th column). Eye fixations in the image-viewing modality are visualized as Gaussian blurred saliency maps with jet colormap. The labels for photo-taking are shown on the left, while those for image-viewing are visualized as color frames (red for ASD and blue for Control).

- **Observation I:** Matched examples in the independent space usually have inconsistent semantic meanings. For example, in rows 1-2 photo sequences with many non-human objects in photo-taking are matched with fixations on human faces in image-viewing.
- **Observation II:** By learning on multi-modality, matched examples in the shared space show consistency in high-level semantic meanings. In rows 1-2 photo sequences with non-human objects are matched with fixations on non-human objects including laptop and coins, while in rows 3-4 photo sequences with many human faces in photo-taking are matched with fixations highly focused on faces in image-viewing.
- **Observation III:** Nearest examples matched in the shared space not only share similar semantic meanings but also have consistent labels for ASD screening. In the independent space, three out of the four rows of examples have inconsistent labels between photo-taking and image-viewing, while in the shared space the labels for the two modalities are the same.

These observations show that, unlike the independent space, the shared space learned by our method is able to bridge the two modalities with high-level semantic concepts. Moreover, the capability of accurately matching samples with consistent labels, *i.e.* Observation III, indicates that our method can correlate two modalities not only with their visual appearance, *e.g.* semantic concepts, but also their predictive labels. As a result, we are able to align features with similar semantic meanings and the same labels (*i.e.* ASD or Control in our context) in the two modalities, allowing different modalities to complement each other by transferring their modal-specific knowledge and thus improving the respective model performance.

Quantitative Evaluation. To further support our findings in the qualitative evaluation, especially Observation

III, we conduct a cross-modal matching and retrieval experiment to quantitatively evaluate the effectiveness of the shared space on correlating examples of consistent labels. Specifically, given the source input from one modality (photo-taking, totally 45 samples are utilized) with a specific predictive label, we compute the accuracy of matching them with input in another modality (image-viewing) that has the same label. Results show that, with our multi-modal distillation method, the Recall@5 (percentage of source input having consistent label with at least 1 out of 5 matched inputs in another modality) is improved from 62.2% (independent space) to 95.6% (shared space), confirming our observation in the qualitative experiments.

5. Conclusion

In this paper, we propose an ASD screening with privileged modality framework, which integrates abundant information from two distinct modalities, *i.e.* photo-taking and image-viewing, and mutually boosts the performance on each of them. Our framework carries three major novelties, including two DNN models for ASD screening on the two modalities and a multi-modal distillation method that distills multi-modal knowledge from a shared space to each modality. It does not require one-to-one pairwise relationship between modalities or the availability of all modalities at testing, providing a general paradigm to take advantage of multiple sources of data in real-world clinical settings. Experimental results show that the proposed models can achieve the new state-of-the-art results, and distilling knowledge across the two modalities further improves their performance by a considerable margin.

Acknowledgements

This work is supported by NSF Grants 1908711 and 1849107.

References

- [1] Anna Anzulewicz, Krzysztof Sobota, and Jonathan T. Delafield-Butt. Toward the autism motor signature : gesture patterns during smart tablet gameplay identify children with autism. *Scientific Reports*, 6, 2016.
- [2] Jon Baio, Lisa Wiggins, Deborah L. Christensen, and 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 Summ* 2018, 67 (No. SS-6):123., 2018.
- [3] Jessica Bradshaw, Amanda Mossman Steiner, Grace Genoux, and Lynn Kern Koegel. Feasibility and effectiveness of very early intervention for infants at-risk for autism spectrum disorder: A systematic review. *Journal of Autism and Developmental Disorders*, 45(3):778–794, 2015.
- [4] Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, July 1997.
- [5] Lluís Castrejón, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2940–2949, 2016.
- [6] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):43:1–43:20, 2017.
- [7] Geraldine Dawson, Andrew N. Meltzoff, Julie Osterling, Julie Rinaldi, and Emily Brown. Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders*, 28(6):479–485, 1998.
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [9] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Zhaohui Che, Yi Fang, Xiaokang Yang, Jesús Gutiérrez, and Patrick Le Callet. A dataset of eye movements for the children with autism spectrum disorder. In *Proceedings of the 10th ACM Multimedia Systems Conference, MMSys '19*, pages 255–260, 2019.
- [10] Nuno C. Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *The European Conference on Computer Vision (ECCV)*, 2018.
- [11] Alex Graves, Abdel rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.
- [12] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
- [14] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [16] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016.
- [17] Ming Jiang and Qi Zhao. Learning visual attention to identify people with autism spectrum disorder. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3287–3296, 2017.
- [18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [19] Seth D. Knig and Elizabeth A. Buffalo. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of Neuroscience Methods*, 227:121 – 131, 2014.
- [20] John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [21] Lauren E. Libero, Thomas P. DeRamus, Adrienne C. Lahti, Gopikrishna Deshpande, and Rajesh K. Kana. Multimodal neuroimaging based classification of autism spectrum disorder using anatomical, neurochemical, and white matter correlates. *Cortex*, 66:46 – 59, 2015.
- [22] Wenbo Liu, Ming Li, and Li Yi. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8):888–898, 2016.
- [23] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [24] Anne Masi, Marilena M. DeMayo, Nicholas Glozier, and Adam J. Guastella. An overview of autism spectrum disorder, heterogeneity and treatment options. *Neuroscience Bulletin*, 33(2), 183193., 2017.
- [25] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [26] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
- [27] James M. Rehg, Agata Rozga, Gregory D. Abowd, and Matthew S. Goodwin. Behavioral imaging and autism. *IEEE Pervasive Computing*, 13(2):84–87, 2014.
- [28] Noah J. Sasson, Jed T. Elison, Lauren M. Turner-Brown, Gabriel S. Dichter, and James W. Bodfish. Brief report: Circumscribed attention in young children with autism. *Journal of Autism and Developmental Disorders*, 41(2):242–247, 2011.

- [29] James W. Tanaka and Andrew Sung. The “eye avoidance” hypothesis of autism face processing. *Journal of Autism and Developmental Disorders*, 46(5):1538–1552, 2016.
- [30] Laurens van der Maaten and Geoffrey Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9 (Nov):2579–2605, 2008.
- [31] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015.
- [32] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544 – 557, 2009. Advances in Neural Networks Research: IJCNN2009.
- [33] Vladimir N. Vapnik. An overview of statistical learning theory. *Trans. Neur. Netw.*, 10(5):988–999, 1999.
- [34] Jun Wang, Qian Wang, Jialin Peng, Dong Nie, Feng Zhao, Minjeong Kim, Han Zhang, Chong-Yaw Wee, Shitong Wang, and Dinggang Shen. Multi-task diagnosis for autism spectrum disorders using multi-modality features: A multi-center study. *Human Brain Mapping*, 38(6):3081–3097, 2017.
- [35] Shuo Wang, Shaojing Fan, Bo Chen, Shabnam Habimi, Lynn K. Paul, Qi Zhao, and Ralph Adolphs. Revealing the world of autism through the lens of a camera. *Current Biology*, 2016.
- [36] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604 – 616, 2015.
- [37] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28, 2014.