ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094684

VITASD: ROBUST VISION TRANSFORMER BASELINES FOR AUTISM SPECTRUM DISORDER FACIAL DIAGNOSIS

Xu Cao[1,*], *Wenqian Ye*[1,*], *Elena Sizikova*[1], *Xue Bai*[2], *Megan Coffee*[3], *Hongwu Zeng*[2], *Jianguo Cao*[2,†]

[1] New York University, New York, USA
[2] Shenzhen Children's Hospital, Shenzhen, China
[3] NYU Grossman School of Medicine, New York, USA
†Corresponding Author.

ABSTRACT

Autism spectrum disorder (ASD) is a lifelong neurodevelopmental disorder with very high prevalence around the world. Research progress in the field of ASD facial analysis in pediatric patients has been hindered due to a lack of well-established baselines. In this paper, we propose the use of the Vision Transformer (ViT) for the computational analysis of pediatric ASD. The presented model, known as ViTASD, distills knowledge from large facial expression datasets and offers model structure transferability. Specifically, ViTASD employs a vanilla ViT to extract features from patients' face images and adopts a lightweight decoder with a Gaussian Process layer to enhance the robustness for ASD analysis. Extensive experiments conducted on standard ASD facial analysis benchmarks show that our method outperforms all of the representative approaches in ASD facial analysis, while the ViTASD-L achieves a new state-of-the-art. Our code and pretrained models are available at `https://github.com/IrohXu/ViTASD`.

***Index Terms*—** Autism Spectrum Disorder, Transfer Learning, Vision Transformer (ViT), Knowledge Distillation (KD)

1. INTRODUCTION

Over the past decade, the significant clinical and scientific value of computer-assisted diagnosis (CAD) based on machine learning has been increasingly recognized. In particular, neural networks and transfer learning offer the benefit of learning compact and fixed dimensional representations from large-scale public datasets and utilizing the resulting representation to finetune models in various fields of medicine. Recent research shows that neural networks can be effective clinical aids for mental illness prevention [1]. However, to date, the progress of neural network approaches applied to the analysis of pediatric autism spectrum disorder (ASD) is limited, in part due to the fact that ASD is a heterogeneous neurodevelopmental disorder with complex cognitive features [2]. As a result, there are substantial difficulties in collecting ASD patient data and designing accurate CAD systems. Considering the high prevalence of ASD children, there is an urgent need for more robust ASD early diagnosis tools in clinical practice.

Quantifiable indices for ASD diagnosis have received much attention [2, 3]. In previous studies, most neural network-based diagnoses of ASD focused on neuroimaging-based approaches [4]. However, neuroimaging data collection is often challenging due to non-cooperation from pediatric patients [5]. Recently, techniques based on behaviour analysis and affective computing have been introduced to address some of these issues. Their key idea is that patients with ASD exhibit altered attention and emotion to specific features of visual information [3]. For ASD children, changes are reflected in facial expressions and eye movement information [6]. Compared with traditional neuroimaging methods, these new methods predict potential ASD risk by directly analyzing a patient's face, eye-tracking (ET) and behavior, all of which have the potential of integrating with other standardized assessments, such as the autism diagnostic observation schedule (ADOS).

Recent studies show that neural network-based methods can successfully distinguish between ASD and non-ASD children, given sufficient and well-annotated facial images [7]. In this work, Hosseini et al. explored different convolutional neural networks (CNN) for facial analysis in ASD children and noted that some models, such as MobileNet [8], can attain an accuracy of 90% in classification of ASD and non-ASD patients. These results imply that neural networks can learn useful facial risk markers of ASD. In other works, Xie et al. [9] and Han et al. [10] designed neural network models using ET data to predict ASD. Their experiments demonstrated that image data (face, ET) collected in vitro is a very promising direction for designing ASD CAD systems.

Existing facial analysis methods for ASD detection rely on CNNs for local feature extraction or as additional eye-tracking object detection modules. On the other hand, ViTs,

*Equal contribution.

This work was supported by Guangdong Province High-level Hospital Construction Project (Shenzhen Children's Hospital, 2021). Xu Cao and Wenqian Ye are IEEE Member.

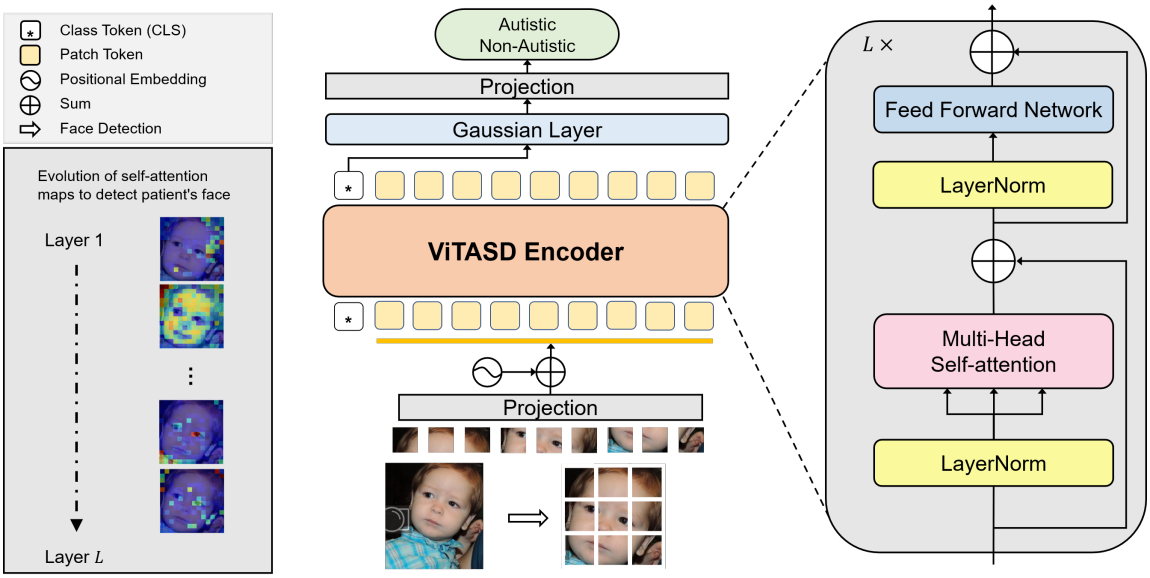

Fig. 1. Overview of the components of the proposed ViTASD model for pediatric autism spectrum disorder (ASD) prediction from facial images. ViTASD consists of a Vision Transformer (ViT) encoder, a Gaussian layer-based decoder and a knowledge distillation (KD) module.

in comparison to CNNs, better capture global context and are less biased toward local textures [11]. However, in small datasets, ViTs often perform worse than CNNs. This limitation motivates us to think from a new direction: how to utilize the strong representation learned from the pretrained ViT in ASD facial analysis? In this work, we propose ViTASD, a novel KD transformer baseline for ASD facial analysis. Our contributions can be summarized as follows.

- We propose ViTASD, a novel model for automatic prediction of the autism spectrum disorder (ASD) in pediatric patients from facial images. ViTASD obtains state of the art performance on the autism spectrum disorder children's dataset [12], a public benchmark.
- We empirically demonstrate capabilities of the ViTASD model: ability to scale model size, support of newest masked autoencoder (MAE) self-supervised learning, knowledge transferability from a large-scale facial expression dataset and large pretrained models.

2. METHODOLOGY

An overview of our proposed framework can be seen in Fig. 1. The goal of ViTASD is to provide a simple yet robust baseline for pediatric ASD detection from facial images. Thus, we aim to keep the original ViT structure [13] without the addition of more complex modules. The decoder of ViTASD is a lightweight multilayer perceptron (MLP) layer with an optional Gaussian layer for Out-of-Distribution (OOD) data points.

2.1. Structure of ViTASD baseline

Given a patient face image $X \in R^{H\times W\times 3}$, ViTASD slices the input image into 16×16 patches via the patch embedding layer. The patches are then flattened to a $K \in R^{(\frac{H\times W}{16\times 16}+1)\times D}$ output with an additional class token. Here, D is the channel dimension. Finally, the tokens are processed by several Transformer blocks, each composed of a multi-head self-attention layer (MHA) and a multi-layer perceptron layer (MLP):

$$K'_{i+1} = MHA_{i+1}((LN(K_i)) + K_i \quad (1)$$

$$K_{i+1} = MLP_{i+1}(LN(K'_{i+1})) + K'_{i+1}, \quad (2)$$

where i is the output of i-th Transformer block; MHA_{i+1} is the multi-head self-attention layer of the $i+1$-th Transformer block; MLP_{i+1} is the multi-layer perceptron of the $i+1$-th Transformer block.

2.2. Properties of ViTASD

Pretraining data flexibility. In contrast to CNN-based methods, ViTASD benefits from the data flexibility from ViT in both transfer learning and representation learning. In our model, we explore data flexibility other than the default settings of ImageNet-21k pretraining. In the ablation study (see Sec. 3.2), we prove that both supervised and self-supervised pretraining using AffectNet can improve the ViTASD's performance.

Model structure transferability. We empirically demonstrate that ViTASD can distill knowledge from larger structures to match performance in smaller ones. To fill the structure gap,

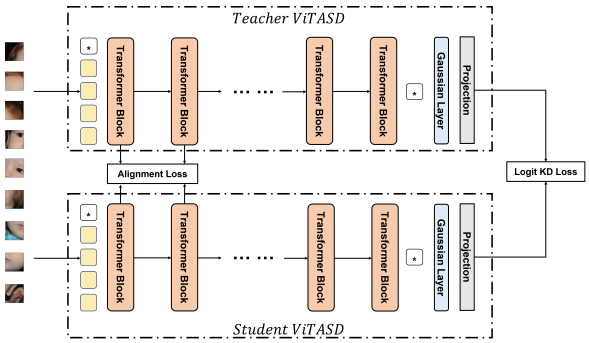

Fig. 2. Knowledge distillation of ViTASD.

we introduce two distillation losses, the alignment loss $\mathcal{L}_{align}$ and the logit KD Loss $\mathcal{L}_{logit}$, to substantially improve performance of the student network (see Fig. 2). $\mathcal{L}_{align}$ aligns the feature distillation on the attention maps of the shallow layers (e.g., layers 0 and 1) using Mean Square Error (MSE), where FC$(\cdot)$ is a linear layer to reshape the $\mathcal{F}^{Teacher}$ to the same dimension as $\mathcal{F}^{Student}$. $\mathcal{L}_{logit}$ ensures the classification logits of the student ViTASD match those of the teacher ViTASD. To summarize, we train the student model with the following total loss:

$$\begin{aligned} \mathcal{L}_{align} &= \text{MSE}(\text{FC}(\mathcal{F}^{Teacher}) - \mathcal{F}^{Student}) \\ \mathcal{L}_{logit} &= \text{MSE}(\text{logit}^{Teacher}(x), \text{logit}^{Student}(x)) \\ \mathcal{L}_{KD} &= \mathcal{L}_{logit} + \alpha\mathcal{L}_{align}, \end{aligned} \tag{3}$$

where α is a hyperparameter to balance the distillation loss.

Finetuning flexibility. We add a Gaussian Process Layer decoder with an RBF kernel to ViTASD to enable finetuning on out-of-domain data. This modification improves the uncertainty representation of the model, hence enhancing its robustness. This layer is implemented as a two-layer network:

$$\text{logits}(x) = \Phi(x)\beta, \quad \Phi(x) = \sqrt{\frac{2}{M}} * \cos(Wx + b) \tag{4}$$

where x is the input, and W and b are frozen weights initialized randomly from Gaussian and uniform distributions, respectively. $\Phi(x)$ are the Random Fourier Features (RFF) [14]. β is a learnable kernel weight similar to that of a Dense layer.

3. EXPERIMENTS

3.1. Implementation details

ViTASD follows the DeiT III [15] architecture and the OOD task setting [16] for ViT, i.e., three augmentations (grayscale, solarization, Gaussian blur) and Cut-Mix, Mix-Up. We use ViT-S [13], ViT-B [13], and ViT-L [13] as encoder backbones and denote the corresponding models as ViTASD-S, ViTASD-B, ViTASD-L. The backbones are initialized with two settings:

Table 1. The performance of ViTASD in different model scales and pretrained settings.

Methods	Params	Pretrained	Accuracy ↑
ResNet50	23.5M	ImageNet-21k	91.00 ± 0.24
ResNet152	60.3M	ImageNet-21k	91.33 ± 0.24
ResNet152	60.3M	AffectNet [18]	89.50 ± 0.24
ViTASD-S	27.1M	ImageNet-21k	91.00 ± 0.41
ViTASD-B	85.8M	ImageNet-21k	92.83 ± 0.24
ViTASD-L	307M	ImageNet-21k	93.17 ± 0.24
ViTASD-L	307M	MAE [17]	94.00 ± 0.41
ViTASD-L	307M	AffectNet [18]	$\mathbf{94.50 \pm 0.23}$

Table 2. Knowledge distillation effect on ViTASD.

Student	Teacher	Teacher Pretrained	Accuracy ↑
ViTASD-S	ViTASD-B	AffectNet [18]	93.50 ± 0.24
ViTASD-B	ViTASD-L	AffectNet [18]	$\mathbf{94.00 \pm 0.24}$

(a) MAE [17] pretrained weights from AffectNet [18] (the largest facial expressions database); (b) general supervised learning pretrained weights from AffectNet [18]. We use the 224×224 input resolution and AdamW optimizer with a learning rate of 1e-4. The α for KD training is 5e-5. All models are trained for 300 epochs with batch size of 128 on 4 NVIDIA A100 GPUs.

We use the largest publicly available facial expression recognition dataset AffectNet [18] (a facial expression dataset with more than 1M images) to pretrain ViTASD in both supervised and self-supervised ways. For self-supervised learning, we adopt MAE [17] by randomly masking 75% patches from the input images and reconstructing those masked patches. The performance of ViTASD is evaluated on the Autism spectrum disorder (ASD) children's dataset [12], which consists of 2,926 images of resolution 224×224 pixels with binary labels (non-autism and autism). The dataset is split into training set (2,526), validation set (200), and test set (200) in its updated version (see [12] for more details). We measure performance using accuracy and area under the receiver operating characteristic (AUROC) on binary labels (non-autism and autism).

3.2. Quantitative evaluation

We report performance comparisons between ViTASD and the state-of-the-art approaches are shown in Table 3. From the results, we make the following observations. (i) A pretrained ViT is significantly better in accuracy and AUROC metrics than any CNN-based methods for ASD facial detection. (ii) The representations learned from the large-scale facial expression dataset (AffectNet) are helpful for transfer learning in ASD. With AffectNet pretraining, the performance

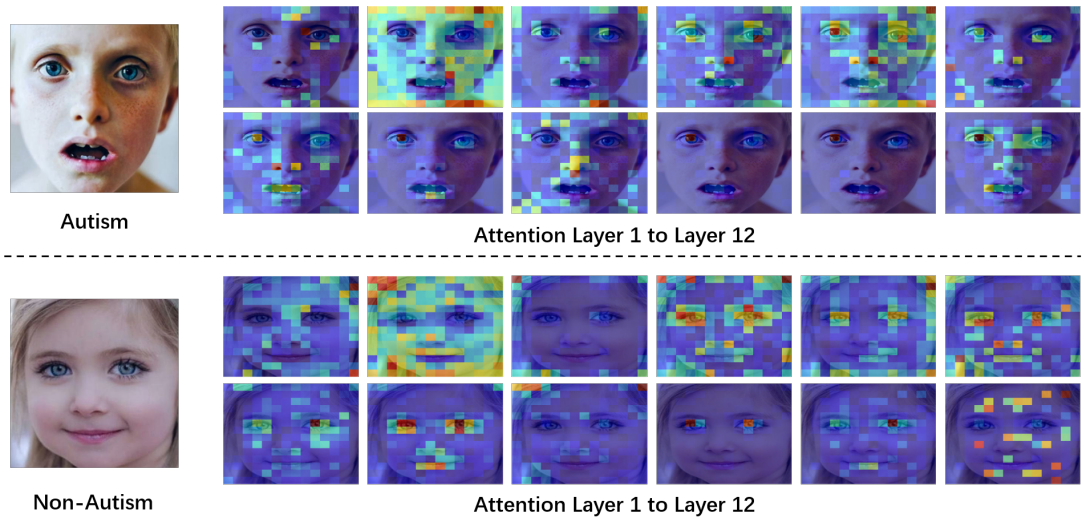

Fig. 3. Visualization of the attention maps between the ASD classification token and all visual tokens in the 12 transformer layers of ViTASD-B, where red and blue represent regions with high and low attention, respectively.

Table 3. Comparison of ViTASD and SOTA methods on the test set of the autism spectrum disorder (ASD) children's dataset [12].

Methods	Backbone	Params	Pretrained	Accuracy↑	AUROC↑
[19, 20]	VGG-19	139M	ImageNet-21k	90.50 ± 0.41	93.65 ± 0.13
[20]	ResNet50	23.5M	ImageNet-21k	91.00 ± 0.23	94.82 ± 0.62
[19, 21, 7]	MobileNetV3	4.2M	ImageNet-21k	91.00 ± 0.23	94.43 ± 0.35
[22]	EfficientNet-B4	17.6M	ImageNet-21k	91.00 ± 0.41	95.13 ± 0.26
[19]	Xception	20.8M	ImageNet-21k	91.33 ± 0.24	95.40 ± 0.16
ViTASD-B	ViT-B	85.8M	ImageNet-21k	92.83 ± 0.24	96.94 ± 0.10
ViTASD-B	ViT-B + knowledge distillation	85.8M	AffectNet [18]	94.00 ± 0.24	97.16 ± 0.48
ViTASD-L	ViT-L	307M	AffectNet [18]	$\mathbf{94.50 \pm 0.23}$	$\mathbf{97.92 \pm 0.12}$

of ViTASD-L further increases to 94.50 accuracy, implying the good knowledge transferability and flexibility of ViTASD.

We also evaluate ViTASD with different ViT backbones and report results in Table 1. We find that accuracy improvement between ResNet-50 and ResNet-152 is significantly lower than between ViTASD-S and ViTASD-B, demonstrating that the ViT model can better learn representation via large facial dataset and transfer into a new ASD facial analysis task. In Table 2, we further investigate the KD performance on ViTASD-B to ViTASD-S and ViTASD-L to ViTASD-B with performance loss of 0.5%. The result illustrates the strong model structure transferability of ViTASD.

3.3. Visualization and interpretability

In order to show the interpretability of the proposed ViTASD, we visualize the attention maps during inference on the test set in Fig. 3. The attention map is the interaction between the classification token and all visual tokens. The attention scores, which showed the color in the attention maps, can be used to understand which areas contribute most to the classification result. For both autism and non-autism children's' face images, the model attends most to the eye region, which is known to be one of the most distinguishable features of the autism children in clinical practice [23].

4. CONCLUSION AND DISCUSSION

In this paper, we propose ViTASD, the first ViT-based baseline for pediatric ASD diagnosis. We have shown that pediatric ASD can be formulated as a facial image classification problem using a ViT, which achieves state-of-the-art performance in both accuracy and AUROC, while generating attention maps consistent with distinguishable ASD features. We hope this work could provide insights to the biomedical imaging and signal processing community for ASD research and inspire further study on exploring the potential of applying explainable ViTs in more facial analysis application tasks.

5. REFERENCES

[1] D. Durstewitz, G. Koppe, and A. Meyer-Lindenberg, "Deep neural networks in psychiatry," *Molecular psychiatry*, 2019.

[2] Xu Cao and Jianguo Cao, "Commentary: Machine learning for autism spectrum disorder diagnosis–challenges and opportunities–a commentary on schulte-rüther et al.(2022)," *Journal of Child Psychology and Psychiatry*, 2023.

[3] R. de Belen, T. Bednarz, A. Sowmya, and D. Del Favero, "Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019," *Translational psychiatry*, 2020.

[4] E. Anagnostou and M. Taylor, "Review of neuroimaging in autism spectrum disorders: what have we learned and where we go from here," *Molecular autism*, 2011.

[5] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, et al., "Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review," *Computers in Biology and Medicine*, 2021.

[6] H. Drimalla, T. Scheffer, N. Landwehr, I. Baskow, S. Roepke, B. Behnia, and I. Dziobek, "Towards the automatic detection of social biomarkers in autism spectrum disorder: Introducing the simulated interaction task (sit)," *NPJ digital medicine*, 2020.

[7] M. Hosseini, M. Beary, A. Hadsell, R. Messersmith, and H. Soltanian-Zadeh, "Deep learning for autism diagnosis and facial analysis in children," *Frontiers in Computational Neuroscience*, 2021.

[8] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[9] J. Xie, L. Wang, P. Webster, Y. Yao, J. Sun, S. Wang, and H. Zhou, "Identifying visual attention features accurately discerning between autism and typically developing: a deep learning framework," *Interdisciplinary Sciences: Computational Life Sciences*, 2022.

[10] J. Han, G. Jiang, G. Ouyang, and X. Li, "A multimodal approach for identifying autism spectrum disorders in children," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[11] M. Naseer, K. Ranasinghe, S. Khan, M. Hayat, F. Shahbaz Khan, and M. Yang, "Intriguing properties of vision transformers," *NeurIPS*, 2021.

[12] G. Piosenka, "Autism spectrum disorder children dataset," `https://drive.google.com/drive/folders/1XQU0pluL0m3TIlXqntano12d68peMb8A`, Accessed: 2022-10-01.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[14] C. Williams and C. Rasmussen, *Gaussian processes for machine learning*, vol. 2, MIT press, Cambridge, MA, 2006.

[15] H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the vit," *arXiv:2204.07118*, 2022.

[16] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *NeurIPS*, 2021.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022.

[18] A. Mollahosseini, B. Hasani, and M. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, 2017.

[19] F. Alsaade and M. Alzahrani, "Classification and detection of autism spectrum disorder based on deep learning algorithms," *Computational Intelligence and Neuroscience*, 2022.

[20] B. Elshoky, E. Younis, A. Ali, and O. Ibrahim, "Comparing automated and non-automated machine learning for autism spectrum disorders classification using facial images," *ETRI Journal*, 2022.

[21] Z. Ahmed, T. Aldhyani, M. Jadhav, M. Alzahrani, M. Alzahrani, M. Althobaiti, F. Alassery, A. Alshaflut, N. Alzahrani, and A. Al-Madani, "Facial features detection system to identify children with autism spectrum disorder: Deep learning models," *Computational and Mathematical Methods in Medicine*, 2022.

[22] K. Mujeeb Rahman and M. Subashini, "Identification of autism in children using static facial features and deep neural networks," *Brain Sciences*, 2022.

[23] Mee-Kyoung Kwon, Adrienne Moore, Cynthia Carter Barnes, Debra Cha, and Karen Pierce, "Typical levels of eye-region fixation in toddlers with autism spectrum disorder across multiple contexts," *Journal of the American Academy of Child & Adolescent Psychiatry*, 2019.