ViTASD: Robust Vision Transformer Baselines for Autism Spectrum Disorder Facial Diagnosis

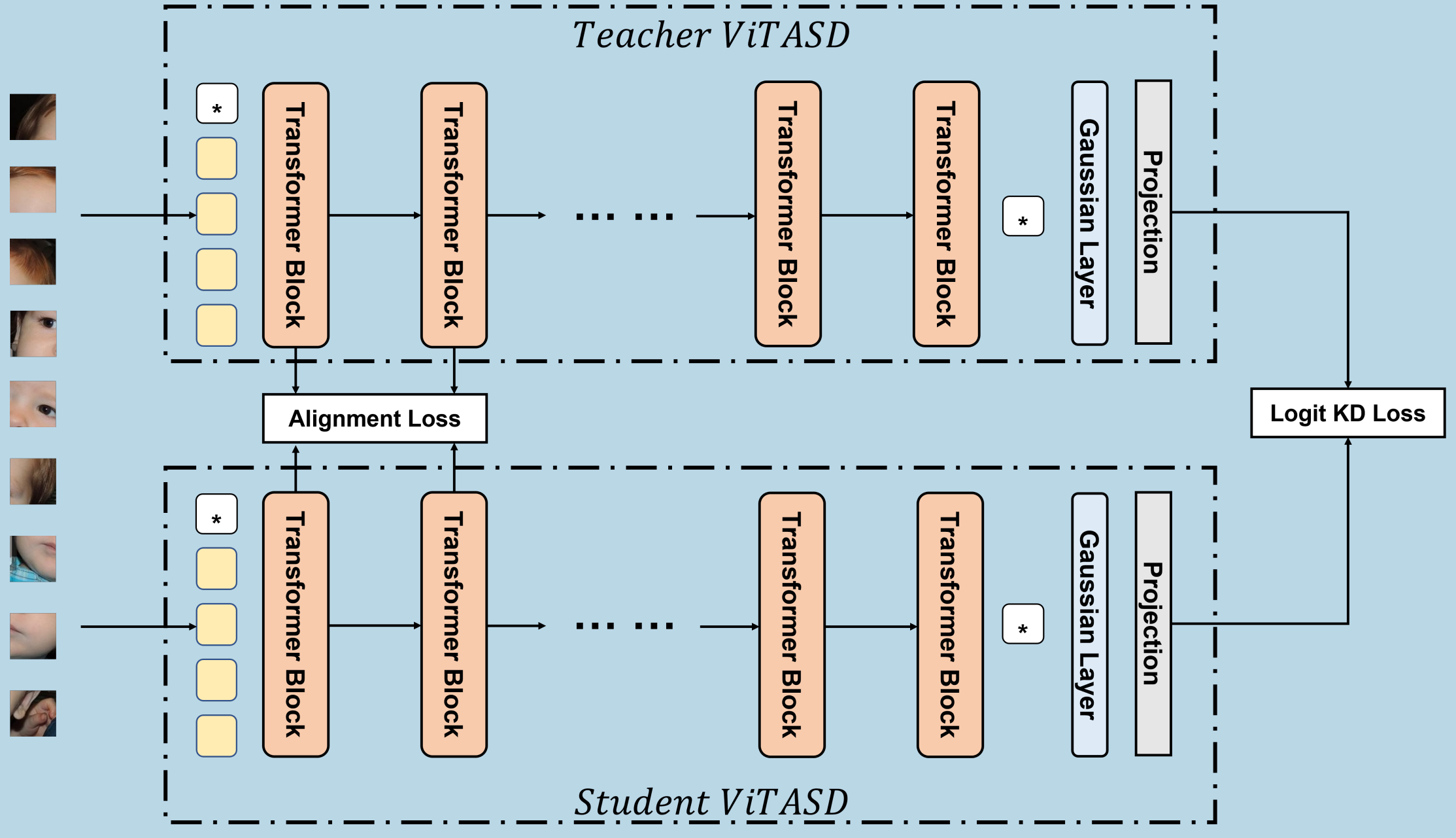
Xu Cao*, Wenqian Ye*, Elena Sizikova, Xue Bai, Megan Coffee, Hongwu Zeng, Jianguo Cao

Challenges

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`.

Model Transferability

We empirically demonstrate that ViTASD can distill knowledge from larger structures to match performance in smaller ones. To fill the structure gap, 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. $\mathcal{L}_{align}$ 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(·) 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.



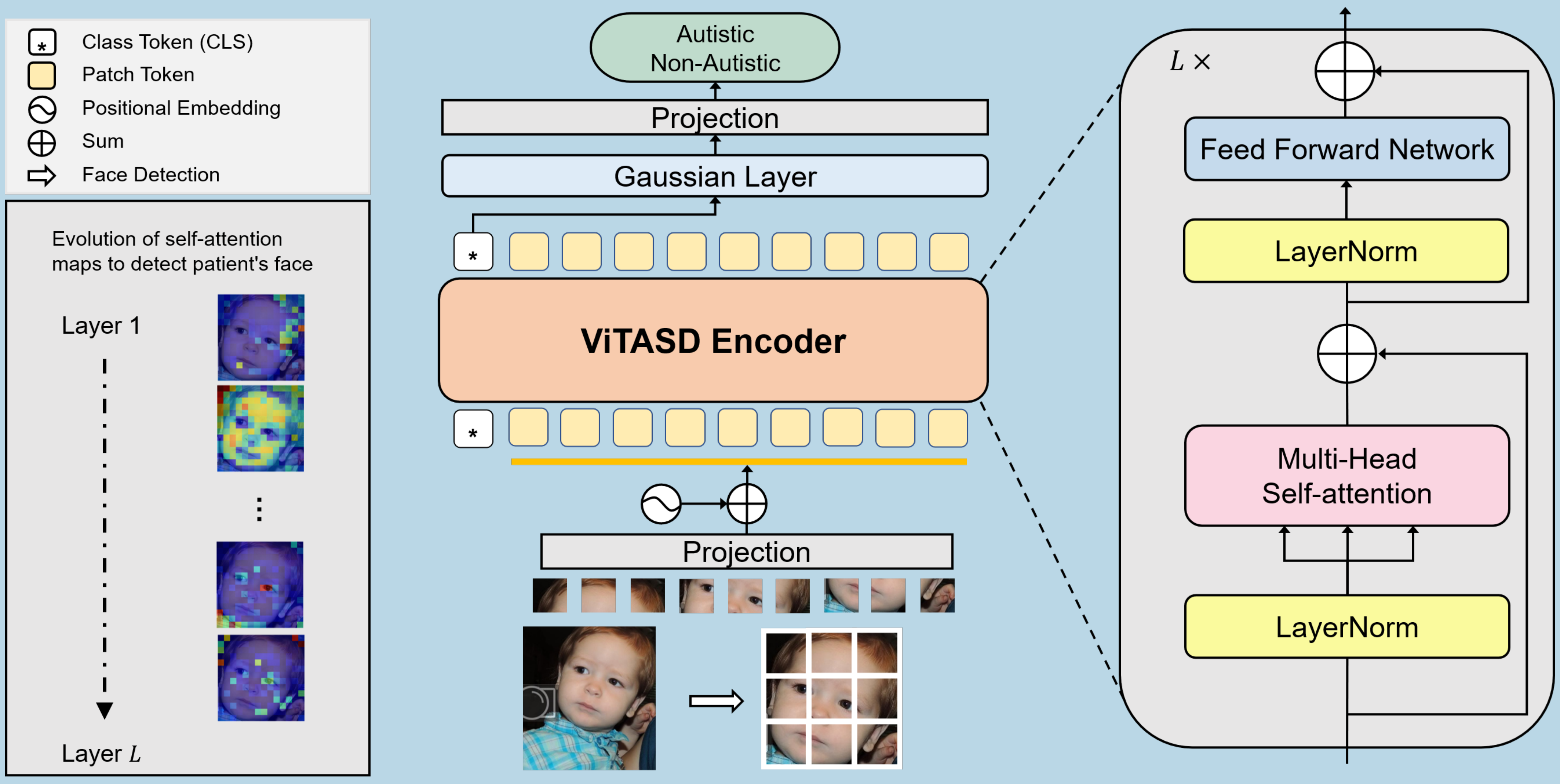
References

[1] Cao, Xu, 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).

Acknowledgements

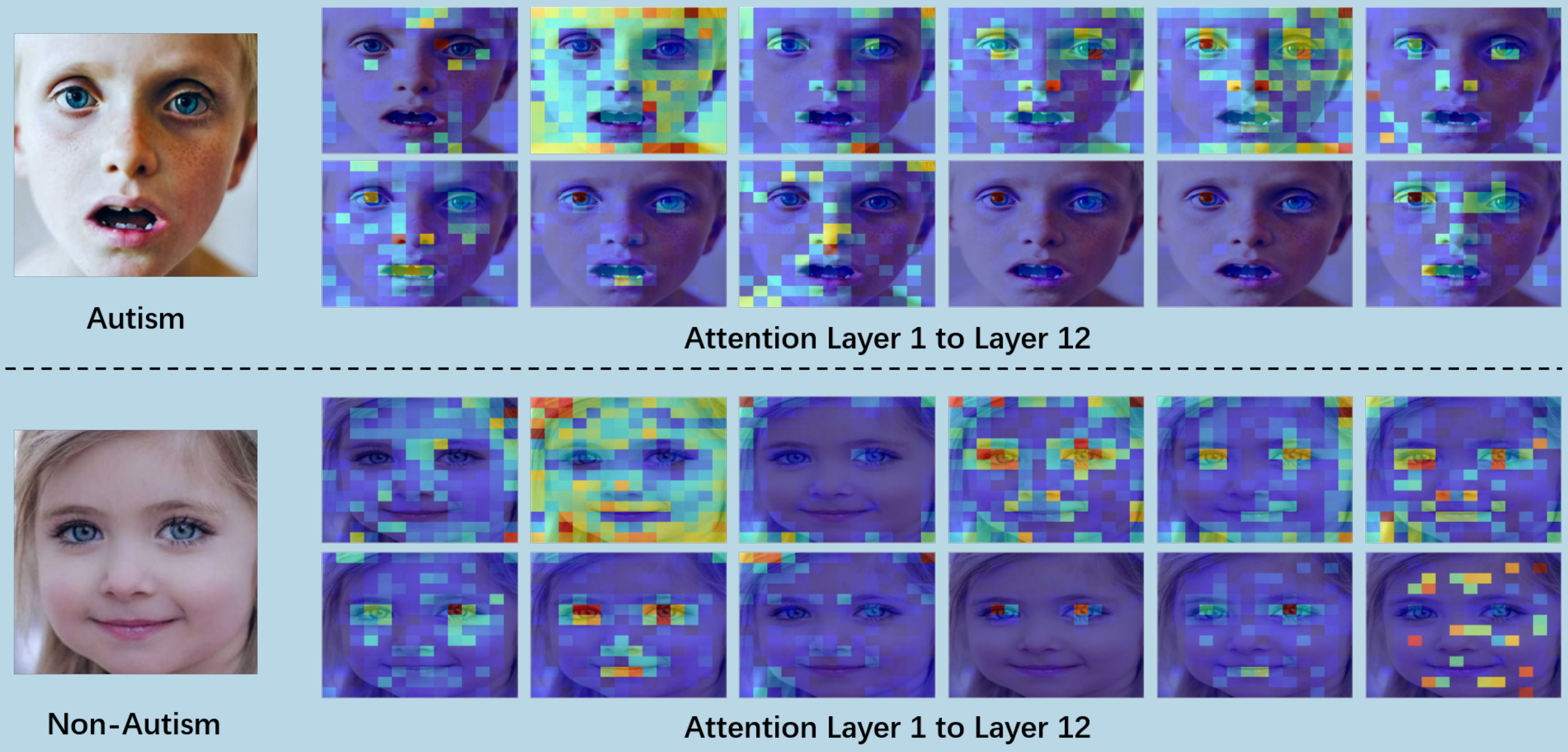
This research was supported by Shenzhen Children's Hospital AI Lab. The corresponding authors are Dr. Jianguo Cao (caojgsz@126.com). The supports are gratefully acknowledged. * denotes equal contribution.

ViTASD Framework



ViTASD model leverages the advantages of Vision Transformers, knowledge distillation, and a Gaussian Process Layer decoder to achieve state-of-the-art performance in automatically predicting ASD from facial images. An overview of our proposed framework can be seen in this figure. 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 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.

Visualization & Interpretability of ViTASD



We report performance comparisons between ViTASD and the state-of-the-art approaches are shown in this table. From the results, we make the following observations. (i) A pre-trained 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 of ViTASD-L further increases to 94.50 accuracy, implying the good knowledge transferability and flexibility of ViTASD.

In order to show the interpretability of the proposed ViTASD, we visualize the attention maps during inference on the test set in this figure. 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 healthy 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.

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]	**94.50 ± 0.23**	**97.92** ± **0.12**