A Lightweight Face Image–Based Auxiliary Detection Model for Autism Spectrum Disorder

Chenkai Liao
School of Computer and Artificial Intelligence,
Hunan University of Technology
Zhuzhou, China
rogersckl@163.com

Wenqiu Zhu
School of Computer and Artificial Intelligence,
Hunan University of Technology
Zhuzhou, China
wenqiu_zhu@126.com

***Abstract*—Early diagnosis of Autism Spectrum Disorder (ASD) plays a crucial role in improving patients' quality of life. In recent years, face image–based ASD detection has attracted increasing attention as an auxiliary diagnostic approach. However, existing lightweight models still show limitations in capturing fine-grained facial features. To address this problem, this paper proposes a lightweight face image–based auxiliary detection model for ASD, termed MN-ASD. First, MobileNetV4-S is selected as the baseline framework. To enhance the model's ability to capture subtle facial details, the FReLU activation function is introduced, which strengthens spatial feature modeling. Second, to further improve the performance and accuracy of the model, the Coordinate Attention (CA) module is incorporated at the final stage. By jointly modeling positional and channel dependencies, the CA mechanism enables the model to better locate key facial regions and enhance global feature extraction. Finally, experimental results demonstrate that the proposed method achieves an accuracy of 93.33% on the Kaggle dataset, with precision, recall, and F1-score all outperforming other mainstream lightweight networks. These results demonstrate the effectiveness of the proposed MN-ASD model in capturing discriminative facial features, while highlighting its strong potential for deployment in resource-constrained and TinyML-oriented environments.**

Keywords—Autism Spectrum Disorder, face image, lightweight, MobileNetV4, Coordinate Attention

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by impairments in social communication and the presence of restricted and repetitive behaviors [1][2].

In recent years, with the rapid development of computer vision technology, researchers have begun to explore the use of objective data such as facial recognition [3], physiological signals [4][5], and eye-tracking [6] to assist in ASD screening. Among these methods, facial recognition has become one of the most important research directions for ASD detection, owing to its convenience of acquisition, strong non-invasiveness, and information-rich nature [7][8].

Early diagnosis and intervention for ASD are of great importance [9]. Studies have shown that children who receive medical care before the age of two achieve higher IQ scores than those diagnosed later in life [10]. Unfortunately, most children are not diagnosed until at least the age of three [11], thereby missing the critical period for language development. Therefore, the development of a simple and effective auxiliary diagnostic tool for ASD is of great practical significance.

Lightweight convolutional neural networks (CNNs) have attracted extensive attention due to their low computational complexity and strong adaptability [12]. However, in ASD facial image detection tasks, existing lightweight models face several limitations. Children with ASD often exhibit subtle and complex facial differences such as gaze avoidance, abnormal movement of the mouth corners, and insufficient coordination of facial expressions [13]. Capturing these features requires not only accurate detection of fine-grained local details—such as eye boundaries and mouth muscle movements—but also attention to the overall coordination of facial expressions. In ASD recognition tasks, models must remain lightweight to adapt to resource-constrained application scenarios, while also possessing sufficient feature modeling capacity to handle the challenges posed by fine-grained facial differences.

Therefore, to address these limitations, this study proposes a lightweight ASD auxiliary detection model (MN-ASD) based on facial images, which integrates the FReLU activation function and the Coordinate Attention (CA) module to enhance feature modeling. The FReLU activation function improves the network's ability to capture fine-grained local features and enhances spatial feature modeling by incorporating a spatially aware conditional branch. The CA module further boosts the model's performance by enabling it to capture both spatial and channel dependencies, making it more capable of locating key facial regions and improving global feature extraction. These innovations allow the proposed model to not only achieve high accuracy but also maintain low computational complexity and parameter quantity, making it suitable for deployment in resource-limited environments. The main innovations of this study are as follows:

- MobileNetV4-S as Backbone: The lightweight network MobileNetV4-S is introduced into ASD facial image detection tasks, providing an efficient and deployable model choice for early auxiliary screening in resource-constrained scenarios.
- Enhanced Feature Modeling: A novel feature modeling strategy is integrated into the model architecture, enabling the network to better capture fine-grained local details in ASD facial images while maintaining lightweight properties.
- Improved Global Representation: Spatial and global dependencies are further incorporated during feature extraction, allowing the model to more accurately

locate key facial regions and enhance overall recognition capability.

II. Literature Review

Facial emotion recognition and facial expression characteristics of individuals with ASD have long been research hotspots in ASD studies. In recent years, deep learning (DL) techniques have shown great potential in the field of ASD detection. Commonly used DL models for facial expression recognition include VGG16, VGG19, and Xception [14][15]. Face image–based DL methods can achieve high diagnostic accuracy across different age groups, offering greater objectivity and reproducibility compared with traditional behavioral observation–based assessments.

Mittal proposed the DL-ASD framework, which utilizes an improved convolutional neural network (I-CNN) to perform multi-label classification on facial images of children aged 1-10 years [16]. Their approach simultaneously identified six types of emotions and four general emotional categories, verifying the feasibility of applying facial features for early ASD screening. However, despite the success of this framework in emotion recognition, it relies heavily on data cleanliness and precise annotations, which may lead to overfitting when trained on small sample datasets, limiting its applicability in resource-constrained environments. Wang employed the EfficientNetB2 model and achieved 91.67% accuracy on the Kaggle dataset, confirming the effectiveness of lightweight architectures in ASD detection tasks [17]. However, although this method improves feature extraction efficiency through the self-attention mechanism, it still falls short in modeling fine-grained local facial features, especially under resource-constrained conditions.

In recent years, several studies on binary classification of ASD based on facial images have employed lightweight networks as backbones to balance efficiency and performance. For example, Beary developed a binary classification model based on MobileNet V1, achieving 94.6% accuracy [18]. Banerjee deployed MobileNetV2 on mobile devices, obtaining an inference latency of only 45 ms, thereby demonstrating the practical value of lightweight models for real-time ASD screening [19]. In addition, Alkahtani combined MobileNetV2 with facial keypoint detection and achieved approximately 92% test accuracy on the Kaggle dataset, further validating the effectiveness of lightweight networks in ASD detection [20]. These methods have indeed demonstrated the advantages of lightweight models in practical applications, but they still face the issue of insufficient fine-grained feature modeling capability. Particularly when dealing with the complex and subtle facial features of individuals with Autism Spectrum Disorder (ASD), existing methods have not fully leveraged the spatial contextual information of the face.

As the latest generation of lightweight networks, MobileNetV4 introduces a unified inverted residual module along with improved activation and convolution combinations, achieving remarkable progress in structural design, computational efficiency, and generalization ability. These advancements make MobileNetV4 particularly well suited for small-sample ASD detection tasks that require deployment on edge devices. Although MobileNetV4 has made significant progress in lightweight design, computational efficiency, and generalization ability, further research is needed to enhance its capability in modeling facial expression feature details and capturing spatial contextual information. Therefore, this study improves upon MobileNetV4 not only preserving its lightweight advantages but also enhancing its ability to model fine-grained features and capture spatial context through the introduction of the FReLU activation function and the Coordinate Attention module, providing a stronger framework for small-sample autism detection tasks.

III. Data Preprocessing

A. Data Sources

This study employed the publicly available Autistic Children Facial Dataset from the Kaggle platform. All experiments were conducted solely for academic research purposes using this publicly available facial image dataset, and no identity recognition or clinical decision-making was involved. To the best of our knowledge, this dataset is currently the only publicly available facial image dataset for autism spectrum disorder research. The images were collected by the publisher from publicly accessible websites. When downloaded, the data is provided in two ways: split into training, testing, and validation versus consolidated [18]. A visual representation of the dataset is shown in Fig. 1.

The dataset supports research on Autism Spectrum Disorder (ASD) screening based on facial images and is divided into two categories according to diagnosis: Autistic (children with ASD) and Non-Autistic (children without ASD).

In total, the dataset contains 2936 images, including 1468 images of autistic children and 1468 images of non-autistic children. The detailed information of this dataset is shown in Table I.

Fig. 1. Autistic children (top) VS Non-autistic children (bottom)

TABLE I. Details of the Autistic Children Facial Dataset

Folder	Trian	Test	Valid
Autistic	1268	150	50
Non_Autistic	1268	150	50

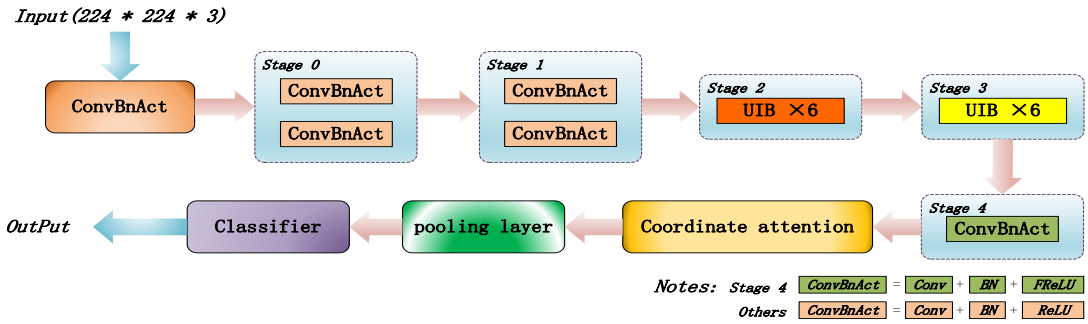

Fig. 2. Structure diagram of MN-ASD model

B. Data augmentation

To mitigate the risk of model overfitting due to the limited size of the training dataset and to further enhance its generalization ability, this study employs an online augmentation strategy during training. This strategy dynamically applies to random transformations each time a training sample is read, without altering the number of samples in the original dataset. Instead, it expands the distribution of effective training data by generating diverse augmented samples in real-time. Therefore, the final dataset size remains the same as the original dataset, while the diversity of the training data is effectively increased.

1) Basic data augmentation

This study systematically combines the following image-level data augmentation methods (applied in the order they appear) to simulate potential visual variations in real-world scenarios:

a) Random Resized Crop: The Random Resized Crop operation is applied, where the input image is randomly cropped and resized to a uniform resolution of 224×224. The area ratio of the cropped region is set in the range of [0.08, 1.0], and the aspect ratio is set between [3/4, 4/3] to enhance the model's robustness to variations in the scale and spatial location of the target object.

b) Random Horizontal Flip: The Random Horizontal Flip operation is applied with a 50% probability to perform horizontal mirroring, which enhances the model's invariance to object orientation.

c) Auto Augment Strategy: The Rand Augment auto-augmentation method is introduced, with the specific configuration set to rand-m9-mstd0.5-inc1. This strategy randomly selects a sequence of augmentation operations from a predefined pool containing geometric transformations and color adjustments. A global augmentation strength of M=9 is applied, with random intensity jittering of 0.5 in each operation, thereby significantly enriching the appearance diversity of the training samples without distorting the image semantics.

d) Color Jitter: In addition to Rand Augment, an independent color jitter with an intensity of 0.4 is applied, randomly perturbing the brightness, contrast, saturation, and hue of the image to simulate lighting and color variations in real imaging environments.

e) Image Normalization: We apply a channel-wise normalization using the standard ImageNet statistics with mean μ = [0.485, 0.456, 0.406] and standard deviation σ = [0.229, 0.224, 0.225]. For each channel c (R, G, B), the normalized value is computed as: $X_c' = \frac{X_C - \mu}{\sigma}$. This normalization is applied consistently in all phases (training, validation, and testing).

f) Random Erasing: After image normalization, a pixel masking operation is applied to random rectangular areas with a probability of 0.1. This regularization technique forces the model not to overly rely on local features, thereby improving its robustness to partial occlusion and noise interference.

2) Label Space Mixing Augmentation

To further enhance the model's generalization and feature discrimination abilities, this study introduces a label-mixing-based strategy in the advanced augmentation stage. Specifically, a combination of Mixup and CutMix is employed, where one of the strategies is randomly selected with a 50% probability for sample synthesis in each training batch. These two methods mix two original samples and their corresponding labels through linear interpolation or region replacement, guiding the model to learn smoother decision boundaries and more robust feature representations.

IV. Model Development and Training Strategies

A. Construction of MN-ASD model

This study adopts the MobileNetV4 family of networks as the basic framework and selects its lightweight variant, MobileNetV4-S, for improvement. As a lightweight convolutional neural network inherited from the MobileNet series, MobileNetV4 [21] combines both efficiency and accuracy. The Small version further reduces the number of parameters and computational cost while maintaining relatively high accuracy, making it more suitable for applications in resource-constrained environments. Therefore, this study employs MobileNetV4-S as the backbone and introduces structural improvements to enhance the model's capability for fine-grained feature modeling.

The improved model, MN-ASD, introduces the FReLU activation function and the Coordinate Attention module. FReLU captures finer-grained spatial features in higher network layers, while the Coordinate Attention module enhances the global dependencies, thereby improving facial recognition accuracy for children with autism. This model retains the lightweight feature extraction capability of MobileNetV4-S, while further strengthening spatial modeling and global dependency capture.

The overall architecture of the improved model is illustrated in Fig. 2, where ConvBnAct denotes the combination of convolution, batch normalization, and activation functions.

The proposed MN-ASD model consists of five stages, in which the feature map resolution gradually decreases from 224×224 at the input to 7×7 at the final stage. Stage 0 to Stage 2 preserve the original MobileNetV4-S structure. In Stage 3 and Stage 4, the original ReLU activation function is replaced with FReLU, thereby enhancing the spatial nonlinear modeling capability of the network. In addition, at the end of Stage 4, the Coordinate Attention (CA) module is introduced to improve the modeling of long-range dependencies and to enhance feature localization. Finally, after feature extraction through the five stages, the classification part of the model is completed through global average pooling, a 1×1 convolution, and a fully connected layer.

B. FReLU Activation function

In the MobileNetV4-S network, ReLU is used as the activation function in Stage 3 and Stage 4 ReLU offers advantages such as computational simplicity and fast convergence; however, it performs nonlinear transformation only at individual points, lacking the ability to model spatial contextual information [22][23]. To overcome this limitation, FReLU extends ReLU by introducing a spatial conditional branch. Its general form can be written as in Eq. (1):

$$y = \max\big(x, T(x)\big) \tag{1}$$

Where x denotes the input feature, and T(x) represents a spatially aware gating term obtained through lightweight convolution. This design allows the activation function to incorporate pixel dependencies within neighborhoods, thereby explicitly introducing spatial awareness.

C. Incorporation of the Coordinate Attention Module

In order to enhance the model's ability to represent and localize key features while maintaining its efficiency, this study integrates the Coordinate Attention (CA) module, proposed by Hou et al [27]. This module is a lightweight attention mechanism specifically designed for mobile networks. Unlike traditional channel attention mechanisms, which excessively compress spatial structural information through global pooling and thereby weaken position-awareness, CA introduces an innovative coordinate information encoding strategy, enabling the joint modeling of channel dependencies and spatial location information.

The module executes two core operations in a feedforward manner: coordinate information embedding and coordinate attention generation. In the coordinate information embedding phase, CA makes critical improvement by replacing global pooling with one-dimensional global average pooling along both the horizontal and vertical spatial directions. Specifically, for the input feature map X, for the c-th channel on height H, pooling is performed along the width W to generate the feature encoding $z_c^h(h)$ in the height direction; similarly, the feature encoding $z_c^w(w)$ is generated in the width direction. This process can be mathematically expressed as shown in Eqs. (2) and (3):

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} \mathrm{x_c}\,(\mathrm{h}, \mathrm{i}) \tag{2}$$

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} \mathrm{x_c}\,(\mathrm{j}, \mathrm{w}) \tag{3}$$

In this way, the two transformations aggregate features along one spatial direction, ensuring that the output maintains precise positional information along the other spatial direction.

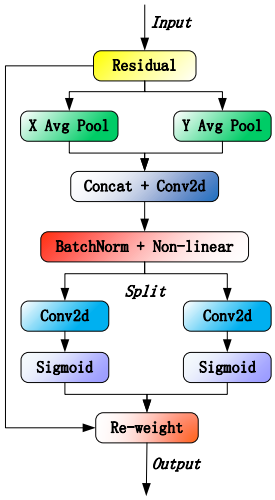

Fig. 3. CA Module Structure

In the coordinate attention generation phase, the feature encodings $z_c^h(h)$ and $z_c^w(w)$ from the coordinate information embedding phase are concatenated and passed through a shared 1×1 convolutional transformation function F_1 for information fusion and dimensionality reduction. Then, an activation function generates an intermediate feature map f. Next, f is split along the spatial dimension into two independent tensors F^h and F^w, which are then transformed back to the same dimensionality as the input channels through 1×1 convolutions F_h and F_w, followed by the application of the Sigmoid function (σ), resulting in direction-aware and position-sensitive attention weights g^h and g^w. Finally, these attention weights are applied to the input feature map X for re-weighting, enhancing the model's focus on target regions. This process can be expressed as shown in Eqs. (4) and (5):

$$g^h = \sigma(F_h(f^h)), g^w = \sigma(F_w(f^w)) \tag{4}$$

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j) \tag{5}$$

Through its unique coordinate information encoding strategy, the CA module significantly improves the network's ability to locate and capture discriminative features, with almost no increase in model complexity. This efficient, lightweight design has been successfully applied in a range of core vision tasks such as image classification [24], object detection [25], and semantic segmentation [26], driving significant performance leaps. The structure of the CA module is shown in Fig. 3.

D. Training Strategy

This study adopted a comprehensive training strategy to ensure effective learning and optimal performance of the proposed model. The Adam W optimizer was selected due to its decoupled weight decay mechanism, which often leads to better generalization compared to the standard Adam optimizer. The initial learning rate was set to 2×10^{-4} after a grid search over {1e-3, 5e-4, 2e-4, 1e-4}, as it provided a stable convergence trajectory on our validation set. Similarly, a weight decay coefficient of 0.025 was chosen to effectively regularize the model without underfitting. To prevent gradient explosion and stabilize the training process, especially in the early layers, adaptive gradient clipping (AGC) was applied with a conservative threshold of 0.02, which we found sufficient to control gradient norms without hindering the learning process. Training hyperparameters are shown in Table II.

TABLE II. HYPERPARAMETERS

Hyperparameter	Value or name
input_size	224×224
batch_size	64
epochs	200
optimizer	AdamW
AdamW_lr	2×10^(-4)
optimizer_weight_decay	0.025
gradient_clipping	AGC
clip_grad	0.02

TABLE III. EFFECTS OF DIFFERENT ACTIVATION FUNCTIONS ON MODEL PERFORMANCE

Activation Functionn	Acc%	Prec%	Rec%	F1%
Relu	92.00	93.15*	90.67	91.89
Leaky ReLU	92.33*	91.50	93.33*	92.41*
PReLu	90.67	89.61	92.00	90.79
FReLu(MN-ASD)	**93.33**	**92.76**	**94.00**	**93.38**

[a.] Bold values indicate the results of the proposed FReLu.

[b.] * Indicates the highest value among the comparison models.

For learning rate scheduling, a cosine annealing schedule combined with a warm-up and cooldown mechanism was employed. This multi-phase strategy provides sufficient exploration during the early training stage, ensures stable convergence in the middle stage, and enables fine-grained parameter optimization in the later stage, thereby maximizing the benefits of the fine-tuning process. The scheduling process is described as follows:

1) Warm-up stage

During the first five epochs, the learning rate was linearly increased from 1×10^{-4} to the maximum learning rate of 2×10^{-4}. This step prevents gradient explosion or excessively fast convergence before the model weights become stable, thereby ensuring smoother feature learning.

2) Cosine annraling stage

After the warm-up stage, the learning rate was periodically decayed according to the following formula, as shown in Eq. (6):

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_{max}}\pi\right)\right) \quad (6)$$

where $\eta_{min} = 1 \times 10^{-5}$, $\eta_{max} = 2 \times 10^{-4}$, and T_{cur} denotes the current training epoch, while the total number of annealing cycles is T_{max}. This strategy gradually decreases the learning rate during the later stage of training, allowing for more fine-grained parameter updates and improving the final stability and generalization ability of the model.

3) Cool down stage

After the cosine annealing stage, an additional 10 epochs of cooldown were introduced, during which the learning rate was fixed at η_{min}. This final phase further stabilized the model weights and reduced the risk of overfitting.

V. ANALYSIS OF EXPERIMENTAL RESULTS

A. Experimental Environment

The experimental environment was configured on a Windows 11 operating system with an Intel(R) Core(TM) i5-14600KF processor and an NVIDIA GeForce RTX 5060 Ti (16 GB) GPU. The software environment was managed using Anaconda, with the deep learning framework built on PyTorch 2.7.1 under Python 3.9.23. GPU acceleration was enabled using CUDA 11.8.

B. Comparative Analysis with Other Activation Functions

Activation functions play a crucial role in convolutional neural networks (CNNs) by introducing nonlinearity, and their selection directly influences the feature representation capability and final classification performance of the model. To investigate the impact of different activation functions on ASD recognition, this study conducted comparative experiments under the condition that all other network structures remained unchanged. Specifically, ReLU, Leaky ReLU, PReLU, and FReLU were selected as activation functions, and their effects on model performance are summarized in Table III.

Among all tested functions, FReLU achieved the best performance, with an accuracy(Acc) of 93.33% and an F1-Score(F1) of 93.38%, while also attaining the highest precision(Prec) and recall(Rec). These results demonstrate that by incorporating a spatial conditional branch, FReLU enhances the modeling of local geometric structures and boundary features, thereby effectively improving recognition performance.

C. Comparative Analysis with Mainstream Lightweight Networks

To verify the effectiveness of the proposed model, several lightweight networks were selected as baselines for comparison, including EdgeNext-S [28], EfficientNetV2-T [29], FastViT-T8 [30], GhostNet [31], and MobileNetV3-S [32]. All models were tested under the same experimental conditions, and the results are reported in Table IV.

Overall, the proposed model outperformed all comparison models in terms of accuracy, precision, recall, and F1-score. In particular, the accuracy of the proposed model reached 93.33%, representing an improvement of approximately 2 percentage points over the relatively strong baseline EfficientNetV2-T (91.33%). Importantly, this superior classification performance was achieved while maintaining both low parameter(Params) count and low computational cost. Compared with EdgeNeXt-S, FastViT-T8, GhostNet, and MobileNetV3-S, the proposed model demonstrated greater stability across all metrics and exhibited clear advantages in both generalization capability and lightweight efficiency.

The performance gains can be attributed to three main factors. First, the introduction of the FReLU activation function in the high-level stages enhances the spatial feature modelling ability of the network. Second, the integration of the CA module at the final stage enables effective capture of channel dependencies and positional information. The improved network architecture significantly enhances feature representation capabilities while maintaining low parameter count (2.64 million) and low computational complexity (0.19 GFLOPs). Overall, the proposed model achieves a favourable balance between classification performance and lightweight efficiency, offering a practical reference for ASD recognition in resource-constrained scenarios.

TABLE IV. COMPARISON OF TEST RESULTS OF DIFFERENT LIGHTWEIGHT NETWORK MODELS

Model	Params	GFLOPs	Acc%	Prec%	Rec%	F1%
EdgeNext-S	5.28	0.96	90.67	87.20	95.33*	91.08
EfficientNetV2-T	12.63	1.91	91.33*	91.33*	91.33	91.33*
FastViT-T8	3.26	0.53	88.00	86.54	90.00	88.24
GhostNet	3.90	0.15	90.67	90.13	91.33	90.73
MobileNetV3-S	1.52*	0.06*	81.00	79.62	83.33	81.43
MN-ASD	**2.64**	**0.19**	**93.33**	**92.76**	**94.00**	**93.38**

[a.] Bold values indicate the results of the proposed MN-ASD model.

[b.] * Indicates the highest value among the comparison models.

TABLE V. COMPARISON OF TEST RESULTS OF DIFFERENT LIGHTWEIGHT NETWORK MODELS

Methods	Param (M)	GFLOPs	Acc%	Prec%	Rec%	F1%
MobileNetV4-S	2.50	0.18	89.33	92.14	86.00	88.97
MobileNetV4-S+CA	2.58	0.19	92.00	93.15	90.67	91.89
MobileNetV4-S+FReLu	2.56	0.19	90.00	88.46	92.00	90.20
MobileNetV4-S+CA+FReLu	2.64	0.19	93.33	92.76	94.00	93.38

Fig. 4 presents the Precision–Recall curves of the proposed MN-ASD model compared with five lightweight benchmark models. As shown in the figure, the MN-ASD model achieves the highest area under the precision–recall curve (AP = 0.9799), outperforming all comparison models. This indicates that the proposed model maintains superior precision across a wide range of recall values, demonstrating its robustness in distinguishing between autistic and non-autistic facial images even under imbalanced data conditions. Compared with other models, MN-ASD exhibits a smoother and more stable curve, especially in the high-recall region, suggesting its stronger capability to correctly identify positive samples without sacrificing precision.

Furthermore, Fig. 5 presents Grad-CAM heatmaps of EdgeNeXt-S, EfficientNetV2-T, FastViT-T8, GhostNet, MobileNetV3-S, and the proposed MN-ASD model, highlighting differences in attention distribution. The lightweight models MobileNetV3-S and GhostNet showed relatively scattered attention during classification, with some heatmaps focusing on background or non-critical regions, thereby weakening the discriminative basis. While EdgeNeXt-S, EfficientNetV2-T, and FastViT-T8 were generally able to capture facial or local features in most samples, their activation regions still exhibited a certain degree of diffusion. By contrast, the proposed MN-ASD model consistently demonstrated stronger discriminative power, with heatmaps showing more concentrated attention on key facial regions and effectively highlighting classification-related details. These results indicate that MN-ASD surpasses other methods in both feature extraction and discriminative capability, exhibiting superior robustness and interpretability.

D. Ablation Study

To validate the effectiveness of the proposed improvements, ablation experiments were conducted by separately introducing the CA module and the FReLU activation function into the baseline MobileNetV4-S network. The experimental results are reported in Table V.

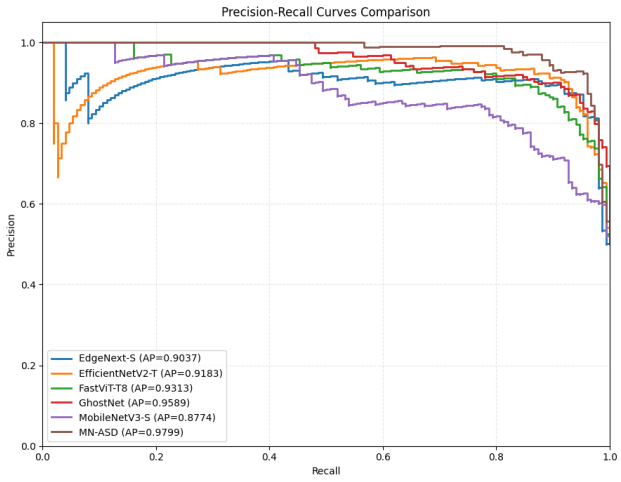

Fig. 4. Precision-Recall Curves Comparison

As shown in Table V, the baseline MobileNetV4-S model achieved an accuracy of 89.33%, with all evaluation metrics remaining at relatively low levels. When only the CA module was incorporated, accuracy increased to 92.00% and F1-Score rose to 91.89%, indicating that CA effectively enhances the modelling of channel dependencies and spatial positional information, thereby improving classification performance. When only FReLU was used in the high-level stages, recall improved to 92.00%, demonstrating that FReLU is advantageous in enhancing spatial feature modelling and boundary awareness. However, the additional spatial branch also altered the feature distribution to some extent, leading to a relative decrease in precision.

When CA and FReLU were combined, the model achieved its best performance. Compared with the baseline MobileNetV4-S, accuracy improved by 4 percentage points, and F1-Score increased by 4.41 percentage points. This indicates that CA and FReLU complement each other, significantly enhancing classification performance on the ASD dataset while preserving the lightweight nature of the model.

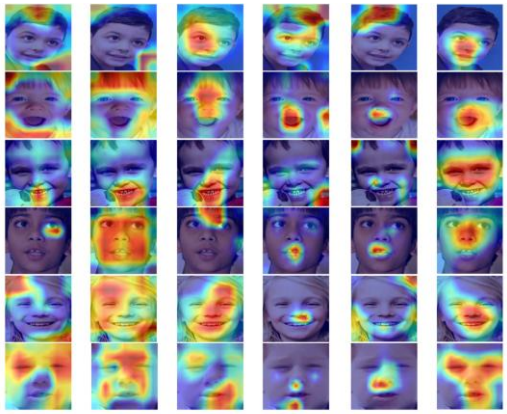

Fig. 5. Grad-CAM (EdgeNext-S(a), EfficientNetV2-T(b), FastViT-T8(c), GhostNet(d), MobileNetV3-S(e), MN-ASD(f))

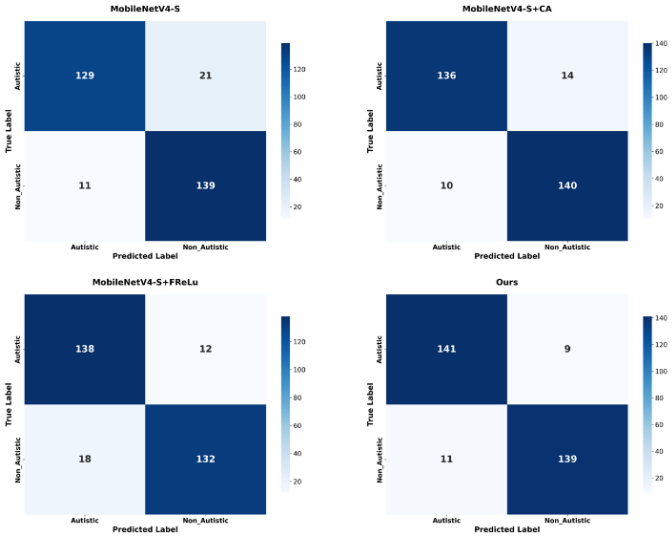

Fig. 6. Confusion matrix

To further analyse the impact of these modules on classification performance, confusion matrices under different experimental settings were plotted, as shown in Fig. 6. For the baseline MobileNetV4-S model, the Autistic category contained a relatively large number of false negatives (FN = 21), and the Non-Autistic category also exhibited a certain number of false positives (FP = 11), resulting in lower overall accuracy. After incorporating the CA module, FN was substantially reduced to 14 and FP decreased to 10, indicating that CA positively contributes to reducing both missed and incorrect diagnoses, thereby improving recall and overall accuracy. With FReLU alone, the number of FN decreased from 21 to 12, and recall improved significantly; however, FP increased to 18, suggesting that while FReLU enhances spatial feature modelling, it may also introduce some misclassifications.

When CA and FReLU were combined, the confusion matrix showed the most favourable outcome: FN decreased to 9, FP was controlled at 11, and the number of correctly classified samples reached the highest values (TP = 141, TN = 139). These results demonstrate that the joint use of CA and FReLU provides complementary advantages.

E. Model Performance Evaluation

To comprehensively evaluate the classification performance of the proposed model, this study employs the ROC curve and AUC with a 95% confidence interval as the core evaluation metrics. As shown in Fig. 8, the model demonstrates exceptional binary classification capability on the test set. The model achieves an AUC of 0.9750 (95% confidence interval: 0.9575–0.9892), indicating its ability to distinguish between positive and negative samples with high accuracy.

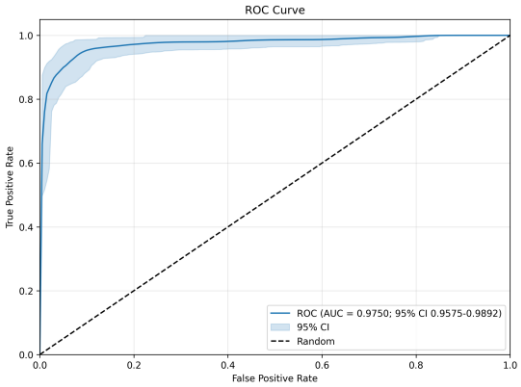

Fig. 7. ROC curve with a 95% confidence interval

VI. Discussion

The proposed MN-ASD model demonstrates excellent performance in the facial image-based auxiliary detection task for Autism Spectrum Disorder (ASD), particularly in terms of accuracy, recall, and F1-score, surpassing existing mainstream lightweight network models. By introducing the FReLU activation function and the Coordinate Attention (CA) module, we effectively enhance the model's ability to capture fine-grained facial details and improve global feature extraction. The model particularly excels at handling the subtle differences in facial features exhibited by children with ASD.

Despite the promising results, there are certain limitations in the experiment. Firstly, the deployment of the model on real-world edge devices, as well as its performance in terms of memory usage and power consumption, remains unverified. In practical applications, resource-constrained devices may encounter performance bottlenecks, making optimization and deployment on such devices a key area for future research. Secondly, although the MN-ASD model achieves notable improvements in facial feature extraction, challenges still exist when dealing with extreme or complex facial expression changes. Therefore, enhancing the robustness of the model under dynamic and varied conditions will remain an important research direction in the future.

VII. Conclusion

This study presents the MN-ASD model, which significantly improves the accuracy and efficiency of facial image-based ASD detection through the incorporation of the FReLU activation function and the Coordinate Attention module. Experimental results demonstrate that the model outperforms other lightweight networks across multiple evaluation metrics while maintaining low computational complexity. Future work will focus on the deployment, memory usage, and power consumption optimization of the model in hardware-constrained environments, as well as further improving its robustness in complex real-world scenarios.

REFERENCES

[1] T. Hirota and B. H. King, "Autism spectrum disorder: a review," JAMA, vol. 329, no. 2, pp. 157–168, 2023, doi: 10.1001/jama.2022.24142.

[2] Arora, M., Jain, C. S., Baru, L. B., Dadi, K., and Raju, B. S. (2024). HyperGALE: ASD classification via hypergraph gated attention with learnable hyperedges. In Proc. Int. Joint Conf. Neural Networks (IJCNN), 1–8. doi: 10.1109/IJCNN60899.2024.10651487.

[3] G. C. Michelassi, H. S. Bortoletti, T. D. Pinheiro, T. Nobayashi, F. R. D. Barros, and R. L. Testa, "Classification of facial images to assist in the diagnosis of autism spectrum disorder," unpublished, Research Square preprint, 2021, doi: 10.21203/rs.3.rs-448184/v1.

[4] M. Kulyabin, P. A. Constable, A. Zhdanov, I. O. Lee, D. A. Thompson, and A. Maier, "Attention to the electroretinogram: gated multilayer perceptron for ASD classification," IEEE Access, vol. 12, pp. 52352–52362, 2024, doi: 10.1109/ACCESS.2024.3386988.

[5] A. Aglinskas, J. K. Hartshorne, and S. Anzellotti, "Contrastive machine learning reveals the structure of neuroanatomical variation within autism," Science, vol. 376, no. 6597, pp. 1070–1074, 2022, doi: 10.1126/science.abm0303.

[6] M. Alcañiz, I. A. Chicchi-Giglioli, L. A. Carrasco-Ribelles, J. Marín-Morales, M. E. Minissi, and G. Teruel-García, "Eye gaze as a biomarker in the recognition of autism spectrum disorder using virtual reality and machine learning: a proof of concept for diagnosis," Autism Research, vol. 15, no. 1, pp. 131–145, 2022, doi: 10.1002/aur.2634.

[7] J. Qiang, D. Wu, H. Du, H. Zhu, S. Chen, and H. Pan, "Review on facial-recognition-based applications in disease diagnosis," Bioengineering, vol. 9, no. 7, p. 273, 2022, doi: 10.3390/bioengineering9070273.

[8] M. K. Yeung, "A systematic review and meta-analysis of facial emotion recognition in autism spectrum disorder: the specificity of deficits and the role of task characteristics," Neuroscience and Biobehavioral Reviews, vol. 133, p. 104518, 2022, doi: 10.1016/j.neubiorev.2021.12.032.

[9] Z. Sherkatghanad, M. Akhondzadeh, S. Salari, M. Zomorodi-Moghadam, M. Abdar, and U. R. Acharya, "Automated detection of autism spectrum disorder using a convolutional neural network," Frontiers in Neuroscience, vol. 13, p. 1325, 2020, doi: 10.3389/fnins.2019.01325.

[10] W. Guthrie, A. M. Wetherby, J. Woods, C. Schatschneider, R. D. Holland, and L. Morgan, "The earlier the better: an RCT of treatment timing effects for toddlers on the autism spectrum," Autism, vol. 27, no. 8, pp. 2295–2309, 2023, doi: 10.1177/13623613231172640.

[11] K. L. Goh, S. Morris, S. Rosalie, C. Foster, T. Falkmer, and T. Tan, "Typically developed adults and adults with autism spectrum disorder classification using centre of pressure measurements," in Proc. 2016 IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Shanghai, China, pp. 844–848, 2016, doi: 10.1109/ICASSP.2016.7471645.

[12] H. I. Liu, M. Galindo, H. Xie, L. K. Wong, H. H. Shuai, and Y. H. Li, "Lightweight deep learning for resource-constrained environments: a survey," ACM Computing Surveys, vol. 56, no. 10, pp. 1–42, 2024, doi: 10.1145/3602105.

[13] M. Kangarani-Farahani, M. A. Malik, and J. G. Zwicker, "Motor impairments in children with autism spectrum disorder: a systematic review and meta-analysis," Journal of Autism and Developmental Disorders, vol. 54, no. 5, pp. 1977–1997, 2024, doi: 10.1007/s10803-023-06487-2.

[14] M. S. Alam, M. M. Rashid, R. Roy, A. R. Faizabadi, K. D. Gupta, and M. M. Ahsan, "Empirical study of autism spectrum disorder diagnosis using facial images by improved transfer learning approach," Bioengineering, vol. 9, p. 273, 2022, doi: 10.3390/bioengineering9100273.

[15] A. Lu and M. Perkowski, "Deep learning approach for screening autism spectrum disorder in children with facial images and analysis of ethnoracial factors in model development and application," Brain Sciences, vol. 11, p. 473, 2021, doi: 10.3390/brainsci11040473.

[16] R. Mittal, V. Malik, and A. Rana, "DL-ASD: a deep learning approach for autism spectrum disorder," in Proc. 5th Int. Conf. Contemp. Comput. Informatics (IC3I), Noida, India, pp. 1–6, 2022, doi: 10.1109/IC3I55659.2022.10051582.

[17] C. Wang and J. Miao, "Deep learning-based recognition of autism using facial datasets," in Proc. Int. Conf. Comput. Vis., Robot., Autom. Eng. (CRAE 2024), SPIE, vol. 13249, pp. 1–10, 2024, doi: 10.1117/12.3042402.

[18] M. Beary, A. Hadsell, R. Messersmith, and M. P. Hosseini, "Diagnosis of autism in children using facial analysis and deep learning," unpublished, arXiv:2008.02890, 2020. [Online]. Available: https://arxiv.org/abs/2008.02890.

[19] A. Banerjee, P. Washington, C. Mutlu, A. Kline, and D. P. Wall, "Training and profiling a pediatric emotion recognition classifier on mobile devices," unpublished, arXiv:2108.11754, 2021. [Online]. Available: https://arxiv.org/abs/2108.11754.

[20] H. Alkahtani, T. H. H. Aldhyani, and M. Y. Alzahrani, "Deep learning algorithms to identify autism spectrum disorder in children-based facial landmarks," Applied Sciences, vol. 13, no. 8, p. 4855, 2023, doi: 10.3390/app13084855.

[21] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, and F. Yang, "MobileNetV4: universal models for the mobile ecosystem," in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland, pp. 78–96, 2024, doi: 10.1007/978-3-031-73661-2_5.

[22] I. Mirzadeh, K. Alizadeh, S. Mehta, C. C. D. Mundo, O. Tuzel, and G. Samei, "Relu strikes back: exploiting activation sparsity in large language models," unpublished, arXiv:2310.04564, 2023. [Online]. Available: https://arxiv.org/abs/2310.04564.

[23] S. Zhang, J. Lu, and H. Zhao, "Deep network approximation: beyond ReLU to diverse activation functions," Journal of Machine Learning Research, vol. 25, no. 35, pp. 1–39, 2024..

[24] J. Ai, Z. Qu, Z. Zhao, Y. Zhang, J. Shi, and H. Yan, "An SAR target classification algorithm based on the central coordinate attention module," IEEE Sensors Journal, vol. 24, no. 2, pp. 1941–1952, 2024, doi: 10.1109/JSEN.2023.3338218.

[25] F. Xie, B. Lin, and Y. Liu, "Research on the coordinate attention mechanism fuse in a YOLOv5 deep learning detector for the SAR ship detection task," Sensors, vol. 22, p. 3370, 2022, doi: 10.3390/s22093370.

[26] T. H. Tsai and Y. W. Tseng, "BiSeNet V3: bilateral segmentation network with coordinate attention for real-time semantic segmentation," Neurocomputing, vol. 532, pp. 33–42, 2023.

[27] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 13713–13722, 2021, doi: 10.1109/CVPR46437.2021.01353.

[28] M. Maaz, A. Shaker, H. Cholakkal, S. Khan, S. W. Zamir, and R. M. Anwer, "Edgenext: efficiently amalgamated CNN-transformer architecture for mobile vision applications," in Proc. Eur. Conf. Comput. Vis. (ECCV), Cham, Switzerland, pp. 3–20, 2022, doi: 10.1007/978-3-031-25082-8_1.

[29] M. Tan and Q. Le, "EfficientNetV2: smaller models and faster training," in Proc. Int. Conf. Mach. Learn. (ICML), PMLR, pp. 10096–10106, 2021.

[30] P. K. A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: a fast hybrid vision transformer using structural reparameterization," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 5785–5795, 2023.

[31] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: more features from cheap operations," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1580–1589, 2020, doi: 10.1109/CVPR42600.2020.00163.

[32] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, and M. Tan, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 1314–1324, 2019, doi: 10.1109/ICCV.2019.00140.