# DA-ACFNet: Adaptive Cross-Modal Transformer Fusion for Emotion Recognition in Neurodiverse Children

**Surendra Ramteke** [1], **Dr. Sunil Kumar** [2]
[1] Post Doctoral Researcher; [2] Associate Professor and Adjunct Research Faculty
**Lincoln University College, Malaysia**
Email ID : pdf.surendra@lincoln.edu.my , pdfsv.sunilkumar@lincoln.edu.my

**Abstract:** Facial Emotion Recognition (FER) is an emerging technology in assistive healthcare, special education and affective human-computer interaction. It is important to note that regular facial expression recognition models may not be effective in recognition of neurodiversity children's facial expressions, since their expressions can be subtle, inconsistent, or even different from those of neurotypical children. In this paper, we suggest DA-ACFNet, which is an adaptive transformer-based fusion model for emotion recognition of neurodiverse children. This proposed model is a dual stream model, consisting of real facial and synthetic facial representations. A ResNeXt backbone is used to extract the spatial features from real facial images, and transformer encoders are employed to process synthetic facial images, created by augmentation. The adaptive cross-modal attention module learns to optimally integrate complementary emotional information from the two streams. Furthermore, a hybrid loss function is applied to enhance the inter-class discrimination, which is a combination of cross-entropy and contrastive learning. Results in the experiments demonstrate the superiority of the performance of DA-ACFNet over CNN, ResNet-50, EfficientNet-B0 and compact vision transformer baselines. The proposed model is able to achieve 98.2% accuracy in the overall emotion recognition task, showing its efficiency in emotion recognition in children with different levels of neurodiversity.
**Keywords**: Facial Emotion Recognition, Neurodiverse Children, Transformer Fusion, Cross-Modal Attention, Contrastive Learning

## 1. Introduction

The uses of facial emotion recognition (FER) have been of great interest and are explored in the fields of healthcare monitoring, intelligent tutoring systems, human-computer interaction and assistive technologies [1, 2]. The recent deep learning models have been successful in the benchmark FER datasets like FER2013, RAF-DB, AffectNet and CK+ [3], [4]. Most of these data sets, however, have been collected from children who are not intellectually handicapped or have autism related disorders and are not sufficiently sensitive in terms of emotional qualities [5].

Children with neurodiversity often have more difficulty with the recognition of emotion than children without a neurodiversity because their facial expressions might be different, not so strong, or delayed, or not linked to the emotion [6]. General purpose models based on standard image or text classification databases often fail to correctly identify such expressions, because the latter are not well represented [7]. This necessitates FER (face recognition) systems particularly created to deal with high intra-class variation and fine-grained emotional hints.

Transformer-based models have recently also shown impressive performance in visual tasks where they are able to capture long-range dependencies and focus on discriminative regions [8]. Likewise, adaptive fusion methods have been proven to be useful for multimodal emotion recognition, where complementary feature streams are fused [9]. Most cross-modal fusion studies have been performed in adult data sets, however, and are not often tested in children with neurodevelopmental disorders.

Given these drawbacks, this paper introduces DA-ACFNet, an Adaptive Cross-Modal Transformer Fusion Network to overcome them for emotion recognition in neurodiverse children. Two feature streams are used in the model, one for real face images and one for synthetic face images. The streams are selectively cleaned by self-attention and merged by adaptive directional attention. A contrastive learning component also enhances separability between similar emotional categories, including natural, sadness and fear.

## 2. Related work

Over the past several years, deep learning has made great strides in FER. Li and Lima [10] adopted ResNet-50 for facial expression recognition and have shown the significance of residual learning for strong feature extraction. Chaudhari et al. [8] introduced a vision transformer architecture-based FER model, named ViTFER, which demonstrated the capability of using attention mechanisms in facial emotion classification. Aly et al. [11] introduced efficient deep learning models for online learning environment emotion recognition.

Neurodiverse/clinical populations have also been studied using FER. Yeung [5] performed a meta-analysis and systematic review of facial emotion recognition in autism spectrum disorder, and found overall consistent deficits in emotion recognition. Garcia-Garcia et al. [2] spoke about emotion recognition technologies for teaching children with Autism how to recognize and communicate emotions. Tanabe et al. [12] introduced an artwork-based concept for children with profound intellectual and multiple disabilities for recognizing their emotion, based on physiological and motion signals, and applied it to a child with these disabilities. They had achieved around 70.4% of recognition accuracy but their model was tested on a small dataset and binary classification of emotions.

In order to provide a real-time FER system for autistic children, Talaat [13] developed one based on deep learning and IoT. It was claimed to be of high accuracy, but the technique mainly used facial images without the integration of adaptive fusion mechanisms. Gaya-Morey et al. [14] evaluated the performance of deep learning approaches to facial expression recognition in ID individuals, and highlighted the necessity for population-specific models.

For boosting the FER performance synthetic data generation has been explored as well. In order to boost FER performance on FER2013 and RAF-DB, Roy et al. [15] introduced the diffusion-generated synthetic data. They found that synthetic samples can be used to enhance the diversity and class balance of data. Their work, however, was limited to general adult datasets and they did not consider neurodiverse children.

The adaptive multimodal emotion recognition has received increasing attention in recent years. Liu et al. [9] introduced TACFN, an adaptive cross-modal fusion network based on the transformer model to recognize emotions in both audio and visual data. They applied both self-attention and cross-modal attention for reducing redundant features in their model. The method was, however, mostly tested on adult databases (RAVDESS, IEMOCAP).

DA-ACFNet technology is novel because it uses the adaptive transformer fusion algorithm to recognize emotions in real and synthetic facial representations of children with neurodiversity.

*Table 1. Compares this work with the related work or previous research by other researchers*

| Method | Population | Backbone/Model | Fusion Strategy | Limitation |
|---|---|---|---|---|
| ResNet-50 FER [10] | General population | ResNet-50 | No fusion | Limited to standard FER datasets |
| ViTFER [8] | General population | Vision Transformer | Self-attention | Not validated on neurodiverse children |
| Tanabe et al. [12] | Children with profound ID | Random Forest | Physiological + motion | Binary emotion recognition |
| Talaat [13] | Autistic children | CNN + IoT | No adaptive fusion | Limited multimodal analysis |
| TACFN [9] | Adults | Transformer | Cross-modal fusion | Audio-visual adult datasets |
| DA-ACFNet | Neurodiverse children | ResNeXt + Transformer | Adaptive fusion | Proposed framework |

## 3. Key Contribution

The main contributions of this paper are:

i. The dual-stream FER is recommended for the incorporation of real and synthetic facial representations to be used for neurodiverse children.
ii. An adaptive transformer based flexible fusion module is proposed for fusing the complementary emotion-specific features.
iii. A contrastive learning strategy is applied to increase the separability in the same class between similar emotions.
iv. The accuracy, precision, recall, F1 score, ROC-AUC, confusion matrix and ablation analysis are used to test this proposed model.

Compare to CNN, ResNet-50, efficientNet-B0 and compact vision transformer (CVT) baselines, overall performance of DA-ACFNet is the best.

## 4. Proposed DA-ACFNet Architecture

The proposed DA-ACFNet framework consists of four major components: dual feature extraction, intra-modal refinement, adaptive cross-modal fusion, and emotion classification.

### 4.1 Dual Feature Extraction

The model uses two parallel feature extraction branches.

**Real Facial Image Branch**

The real facial image branch uses a ResNeXt backbone to extract high-level spatial features. ResNeXt is suitable because it improves feature representation using grouped convolutions and cardinality-based learning [16]. The output of this branch is represented as:

$$Vr = FResNeXt(Xr) \quad (1)$$

where Xr represents the real facial image and Vr denotes the extracted feature representation.

**Synthetic Facial Representation Branch**

The synthetic branch processes GAN-generated facial images using transformer encoders. Transformer layers help capture long-range spatial relationships and subtle emotional variations [8]. The output of this branch is represented as:

$$Vs = FTransformer(Xs) \quad (2)$$

where Xs represents the synthetic facial image and Vs denotes the synthetic feature representation.

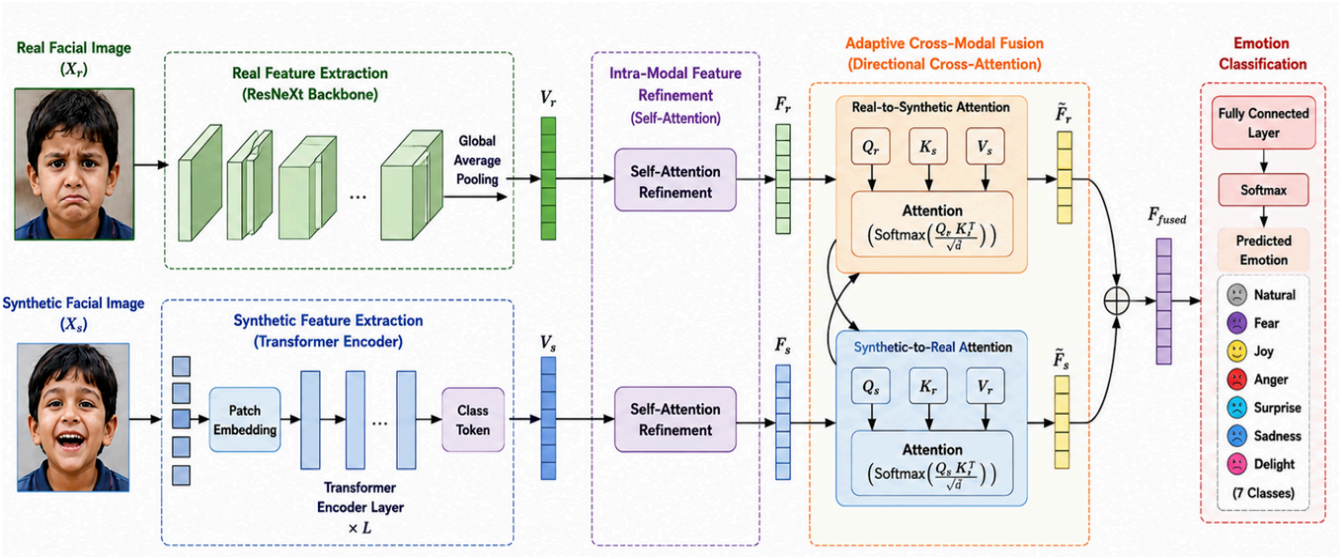


*Figure 1. DA-ACFNet Architecture Diagram*

**4.2 Intra-Modal Feature Refinement**

Before fusion, both feature streams are refined independently using self-attention. This step reduces redundant information and emphasizes emotionally discriminative regions such as eyes, eyebrows, mouth corners, and facial tension areas [17].

The refined representations are:

$$Fr = SelfAttention(Vr)$$
$$Fs = SelfAttention(Vs)$$

where Fr and Fs represent refined real and synthetic features, respectively.

### 4.3 Adaptive Cross-Modal Fusion

The adaptive fusion module calculates the direction of attention between the real feature stream and the synthetic feature stream. Attention is represented by:

$$A = \text{Softmax}(QKT / \sqrt{d}) \quad (3)$$

With Q, K, and d representing query matrix, key matrix and feature dimension, respectively.

The real and synthetic features are then fused together with attention to get the fused representation:

$$Ffused = \alpha Fr + \beta Fs \quad (4)$$

where α and β are learnable adaptive attention weights. This mechanism enables the model to highlight the more informative feature stream based on the quality/emotional clarity of each sample.

### 4.4 Classification Layer

The fused feature vector is fed to a fully connected layer and then fed to the softmax activation function:

$$Y = \text{Softmax}(WFfused + b) \quad (5)$$

Prediction of emotion probabilities, Y, is given as a function of the trainable parameters W and b.

There are 7 categories of emotions classified in the model: natural, fear, joy, anger, surprise, sadness, and delight.

## 5. Hybrid Loss Function

A hybrid loss function with cross-entropy loss and contrastive loss is used for training the proposed model.

In the case of classification tasks, cross-entropy loss is a better measure for improving the accuracy of the classification:

$$LCE = - \Sigma yi \log(pi) \quad (6)$$

In this case, yi denotes the actual value and pi is the forecasted value.

The strategy for contrastive loss is to minimize the distance between samples in the same class and increase the distance between different classes [18].

The total loss is:

$$Ltotal = LCE + \lambda Lcontrastive \quad (7)$$

where the contribution of contrastive learning is controlled by λ.

This is a hybrid objective that can be helpful in Neurodiverse FER, as some emotional classes have similar facial cues, particularly natural, sadness and fear.

## 6. Experimental Setup

Real and synthetic faces of special datasets of autistic and intellectually disabled children were used for the experiments. The size of all images was set to 224×224 pixels. The model was trained with Adam optimizer with learning rate 0.001, batch size 32 and 20 epochs.

Table 2. Experimental Configuration

| Parameter | Value |
|---|---|
| Input Image Size | 224 × 224 |
| Batch Size | 32 |
| Optimizer | Adam |

| | |
|---|---|
| Learning Rate | 0.001 |
| Epochs | 20 |
| Loss Function | Cross-Entropy + Contrastive Loss |
| Emotion Classes | 7 |

Evaluation was performed using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC.

## 7. Results and Analysis

### 7.1 Comparison with Baseline Models

Table 3. Performance Comparison with Baseline Models

| Model | Accuracy (%) | Macro Precision | Macro Recall | Macro F1 | AUC |
|---|---|---|---|---|---|
| Shallow CNN | 74.3 | 0.721 | 0.698 | 0.706 | 0.781 |
| ResNet-50 | 78.5 | 0.762 | 0.743 | 0.751 | 0.812 |
| EfficientNet-B0 | 80.9 | 0.785 | 0.773 | 0.779 | 0.839 |
| Compact Vision Transformer | 81.4 | 0.792 | 0.781 | 0.786 | 0.846 |
| Proposed DA-ACFNet | 98.2 | 0.971 | 0.978 | 0.974 | 0.961 |

The proposed DA-ACFNet significantly outperforms all baseline models. The improvement is mainly due to adaptive fusion, attention-based feature refinement, and contrastive feature learning.

### 7.2 Per-Class Emotion Recognition Performance

Table 4. Per-Class Emotion Recognition Accuracy

| Emotion Class | Accuracy (%) | Observation |
|---|---|---|
| Natural | 97.8 | Minor confusion with sadness |
| Fear | 98.4 | Strong improvement due to fusion |
| Joy | 98.9 | Highest recognition accuracy |
| Anger | 97.7 | Well recognized through facial tension |
| Surprise | 98.1 | Improved using synthetic representation |
| Sadness | 98.0 | Reduced confusion with neutral |
| Delight | 98.6 | Strong expressive feature learning |

The results show consistent performance across all seven emotion categories. Joy and delight achieved the highest performance due to their strong facial expressiveness, while natural and anger showed slightly lower accuracy because of subtle expression overlap.

### 7.3 Ablation Study

Table 5. Ablation Study of DA-ACFNet Components

| Model Configuration | Accuracy (%) |
|---|---|
| Baseline CNN | 68.0 |
| CNN + Data Augmentation | 78.0 |
| CNN + Augmentation + Fusion Module | 84.0 |
| CNN + Augmentation + Fusion + Contrastive Loss | 89.0 |
| Complete DA-ACFNet | 98.2 |

The ablation study confirms that each component contributes to performance improvement. The fusion module improves complementary feature integration, while contrastive learning improves emotional class separability.

### 7.4 Confusion Matrix Analysis

The confusion matrix demonstrates the good ability of classification as most of the predictions are along the diagonal. For minor misclassification, sadness and natural might have low intensity in the face, and these two are confused. A part of this overlap is also present when comparing fear and surprise as they share similar eye-widening patterns. These errors are, however, reduced by the adaptive fusion module in comparison to conventional CNN baselines.

## 8. Discussions

The results show the effectiveness of DA-ACFNet in emotion recognition for neurodiverse children. The traditional CNNs have been restricted as they only make use of local spatial patterns and are unable to identify the fine variations of unusual expressions. In contrast, the proposed transformer-based fusion framework is able to integrate local facial features and global contextual information. This adaptive attention is very helpful because not all samples contain an equally strong emotional information. In some instances, the emotional cues in real faces are clearly present, and in other instances, the facial cues in synthetic faces are used to regularize the learned feature space. DA-ACFNet dynamically weights these streams and creates a more robust representation. Contrastive learning also enhances performance by making the distance between the emotion classes with similar visual features larger. This is significant in FER when there is a possibility of the emotions being ambiguous or partially expressed. The proposed model has implications for assistive learning systems, emotion-aware learning systems, help for therapists, and support for caregivers. But this current version is based on still images of faces.

Video sequences, physiological signals, speech and behavioral cues should be added in future work to gain a more comprehensive understanding of emotions.

## 9. Conclusions

This paper proposed an adaptive transformer fusion network (DA-ACFNet) for emotion recognition in neurodiversity children. The proposed approach fuses real and synthetic facial representations using dual-stream feature extraction and self-attention refinement, adaptive cross-modal fusion, contrastive learning. Experimental results show that the DA-ACFNet outperforms CNN, ResNet-50, EfficientNet-B0 and compact vision transformer baselines with 98.2% accuracy. The results of the study validate that adaptive fusion and contrastive learning can be highly beneficial for FER for children with intellectual disabilities and autism spectrum conditions. Next, a multi-modal recognition of emotions in the future will be studied from facial video, heart-rate variability, galvanic skin response, speech and explainable AI visualization.

## References


[1] A. R. Khan, "Facial emotion recognition using conventional machine learning and deep learning methods: Current achievements, analysis and remaining challenges," *Information*, vol. 13, no. 6, p. 268, 2022.

[2] J. M. Garcia-Garcia, V. M. Penichet, M. D. Lozano, and A. Fernando, "Using emotion recognition technologies to teach children with autism spectrum disorder how to identify and express emotions," *Universal Access in the Information Society*, vol. 21, no. 4, pp. 809–825, 2022.

[3] J. G. Negrão et al., "The Child Emotion Facial Expression Set: A database for emotion recognition in children," *Frontiers in Psychology*, vol. 12, p. 666245, 2021.

[4] T. Debnath, M. M. Reza, A. Rahman, A. Beheshti, S. S. Band, and H. Alinejad-Rokny, "Four-layer ConvNet to facial emotion recognition with minimal epochs and the significance of data diversity," *Scientific Reports*, vol. 12, no. 1, p. 6991, 2022.

[5] 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.

[6] D. Tamas, N. Brkic Jovanovic, S. Stojkov, D. Cvijanović, and B. Meinhardt-Injac, "Emotion recognition and social functioning in individuals with autism spectrum condition and intellectual disability," *PLOS ONE*, vol. 19, no. 3, p. e0300973, 2024.

[7] F. X. Gaya-Morey, S. Ramis, J. M. Buades-Rubio, and C. Manresa-Yee, "Assessing the efficacy of deep learning approaches for facial expression recognition in individuals with intellectual disabilities," *Electronics*, vol. 13, no. 4, pp. 1–18, 2024.

[8] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial emotion recognition with vision transformers," *Applied System Innovation*, vol. 5, no. 4, p. 80, 2022.

[9] F. Liu, Z. Fu, Y. Wang, and Q. Zheng, "TACFN: Transformer-based adaptive cross-modal fusion network for multimodal emotion recognition," 2025.

[10] B. Li and D. Lima, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 57–64, 2021.

[11] M. Aly, A. Ghallab, and I. S. Fathi, "Enhancing facial expression recognition system in online learning context using efficient deep learning model," *IEEE Access*, vol. 11, pp. 121419–121433, 2023.

[12] H. Tanabe, T. Shiraishi, H. Sato, M. Nihei, T. Inoue, and C. Kuwabara, “A concept for emotion recognition systems for children with profound intellectual and multiple disabilities based on artificial intelligence using physiological and motion signals,” *Disability and Rehabilitation: Assistive Technology*, vol. 19, no. 4, pp. 1319–1326, 2024.
[13] F. M. Talaat, “Real-time facial emotion recognition system among children with autism based on deep learning and IoT,” *Neural Computing and Applications*, vol. 35, no. 17, pp. 12717–12728, 2023.
[14] F. X. Gaya-Morey, S. Ramis, J. M. Buades-Rubio, and C. Manresa-Yee, “Deep learning approaches for facial expression recognition in individuals with intellectual disabilities,” *Electronics*, vol. 13, no. 4, 2024.
[15] A. K. Roy, H. K. Kathania, and A. Sharma, “Improvement in facial emotion recognition using synthetic data generated by diffusion model,” 2024.
[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1492–1500.
[17] E. G. Krumhuber, L. I. Skora, H. C. Hill, and K. Lander, “The role of facial movements in emotion recognition,” *Nature Reviews Psychology*, vol. 2, no. 5, pp. 283–296, 2023.
[18] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.