# Two Phase Transfer Learning for Facial Expression Recognition Using ResNet34 with CBAM Attention

B M Rauf
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
b.m.rauf@g.bracu.ac.bd

Anupam Sen Sagor
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
anupam.sen.sagor@g.bracu.ac.bd

Azmari Sultana
*Department of Computer Science and Engineering*
*BRAC University*
Dhaka, Bangladesh
azmari.sultana@g.bracu.ac.bd

***Abstract*—Facial expression recognition (FER) plays a key role in computer applications in human-computer interaction, mental health, and affective computing. However, the existing methods give rise to severe challenges in inter domain generalization, particularly when the RGB trained models are to be utilized in grayscale images. Most of the state-of-the-art methods are either single domain or they do not sufficiently address domain shift between color based and grayscale representations. To address this shortcoming, we propose a two stage transfer learning framework on the basis of integrating ResNet34 with channel and spatial attention models based on Convolutional Block Attention Module (CBAM). To be able to learn strong features of the face, we pre train on Balanced AffectNet (8 emotion classes, RGB) then finetune on FER2013 (7 classes, grayscale) at varying learning rates and augmented with differentiated learning rates and augmentation plans such as RandomErasing to conquer occlusion vulnerabilities. We apply our model on the FER2013 benchmark of 7,178 test samples and achieve 68 percent accuracy with macro F1-score of 0.65, which also demonstrates that our model can be used to adapt between RGB and grayscale domain. It shows that the proposed model is potentially applicable in real-world facial emotion recognition applications where changes in image modality cannot be avoided.**

***Index Terms*—Facial Expression Recognition, Transfer Learning, ResNet34, CBAM Attention, Domain Adaptation**

## I. Introduction

Facial expression recognition (FER) is a branch of machine interaction that helps machines to decode human emotions and is essential in human-computer interface, mental health, educational technology, and surveillance devices [1]. FER is a problem which is usually defined as a multi-class classification problem, where discrete emotions (e.g., Happy, Sad, Anger) are predicted based on facial images. This paper concentrates on in-the-wild, static image-based FER under cross domain transfer between RGB training data to grayscale deployment environment. The low-resolution, noisy and imbalanced grayscale images used in the FER2013 benchmark is a particularly challenging test case [2].

Early FER was based on manual features like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), which were not very generalized [1]. Deep CNNs, which allowed automatic feature learning, attention mechanisms further enhanced performance by focusing on emotion relevant regions on face: Wang et al. in their publication stated that Global Attention yielded 3.8-6.6% improvement in YOLOv8 [1]. Additionally, Alshammari and Alshammari [3] utilized YOLOv8 to demonstrate the effectiveness of advanced CNN architectures for emotion recognition, whereas Ma et al. achieved 73.9% accuracy on the FER2013 with dual channel attention: SE and ECA [4]. Roy et al. [5] further advanced this by proposing ResEmoteNet, which integrates Squeeze-and-Excitation (SE) blocks with residual networks to reduce loss. CBAM [6] has demonstrated good success in vision tasks in both channel and spatial attention. Even with such advances, the majority of them deal with a single domain and do not support the RGB-to-grayscale domain shift. To overcome this point, we introduce a two stage transfer learning system that pre trains a CBAM augmented ResNet34 on Balanced AffectNet and fine tunes it on FER2013. This paper shows that attention-directed transfer learning is an effective method to reduce the RGB to grayscale gap, improve generalization and increased face emotion recognition strength. Our main contributions are:

- Two phase transfer learning pipeline that allows successful RGB to grayscale domain transfer to FER.
- Adaptation of CBAM inspired channel and spatial attention systems into the ResNet34 to emphasize emotion-related face areas.
- An alternative fine tuning learning rate method which balances feature preservation and adaptation at network levels.
- Augmented extensively such as RandomErasing that enhances resistance to occlusion and variation.
- Statistical result to define emotion pairs that are highly confusing, and give feedback on how to make specific improvements.

## II. Related Work

Early FER used handcrafted features (HOG, LBP) which were not good at generalization [1]. CNNs made it possible to extract features automatically. Recent studies investigate attention mechanisms: Wang et al. [1] have obtained 3.8-6.6% gain with Global Attention in YOLOv8. Comparative studies by Kadhim et al. [7] have also analyzed the performance trade-offs between detection models like YOLOv9 and classification models like ResNet50. Ma et al. [4] have introduced dual-channel attention (SE + ECA) in YOLO11-AE, which attains 73.9% on FER2013. The CBAM module [6] is applied sequentially to get the ability to focus on ”what” (channel attention) and ”where” (spatial attention), rather than single-dataset YOLO methods [8]–[10] that do not take into account cross-domain transfer learning. Nathani [11] similarly explored transfer learning using modified VGG16 models, highlighting the specific difficulties in generalizing between FER2013 and AffectNet. Our work integrates spatial attention into ResNet34, addressing cross-domain transfer learning from RGB to grayscale datasets. This aligns with findings by Atymtayeva et al. [12], who demonstrated the superiority of training on AffectNet over FER2013 for Deep CNN model selection.

## III. Methodology

### A. Architecture

We use ResNet34 with attention based on CBAM. ResNet34 balances provide superior capability to resist overfitting compared to deeper variants on FER datasets.

**ResNet34 Backbone:** Skip connections enable gradient flow:

$$\mathbf{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}$$

**Channel Attention:** Models inter-channel dependencies:

$$M_c = \sigma\Big(W_1(\text{ReLU}(W_0(\text{AvgPool}(\mathbf{F}) + \text{MaxPool}(\mathbf{F}))))\Big)$$

with reduction ratio = 16. Output:

$$\mathbf{F}' = M_c \odot \mathbf{F}$$

**Spatial Attention:** Identifies emotion-relevant regions:

$$M_s = \sigma\Big(\text{Conv}_{7\times7}([\text{AvgPool}_c(\mathbf{F}'); \text{MaxPool}_c(\mathbf{F}')])\Big)$$

Final output:

$$\mathbf{F}'' = M_s \odot \mathbf{F}'$$

**Classifier:** Two-layer architecture with dropout: `nn.Sequential(nn.Linear(512, 256), nn.ReLU(), nn.Dropout(0.3), nn.Linear(256, num_classes))`

### B. Phase 1: AffectNet Pre-training

**Dataset:** Balanced AffectNet [13] with 8 emotion classes (Anger, Contempt, Disgust, Fear, Happy, Neutral, Sad, Surprise). Training: 22,900, Validation: 5,700 RGB images from web sources.

Fig. 1 demonstrates that the distribution is balanced between 8 emotion categories with about 2,800-2,900 samples in each category so that there is no bias in the imbalance of classes during the pre-training stage.

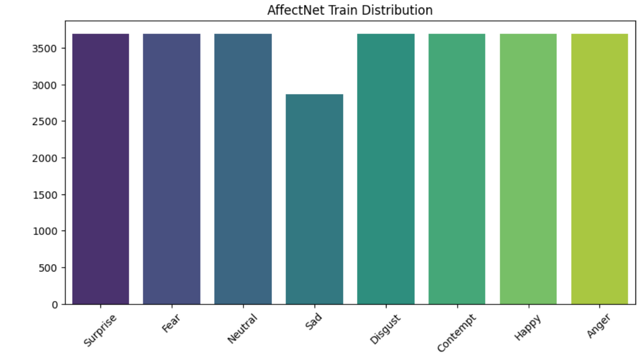

Fig. 1. Balanced AffectNet class distribution.

**Augmentation:** RandomHorizontalFlip (p=0.5), RandomRotation (±8°), ColorJitter (brightness=0.15, contrast=0.15).

**Training:** Batch=32, Adam(lr=0.0001, weight_decay=1e-4), CrossEntropyLoss with class weighting label smoothing (0.1), ReduceLROnPlateau(factor=0.5, patience=3), epochs = 15, mixed-precision training.

Fig. 2 shows the stable convergence of 15 epochs with the training and validation curves on the increase, but no overfitting is observed, which means that the generalizable RGB facial features are learned and can be transferred to FER2013.

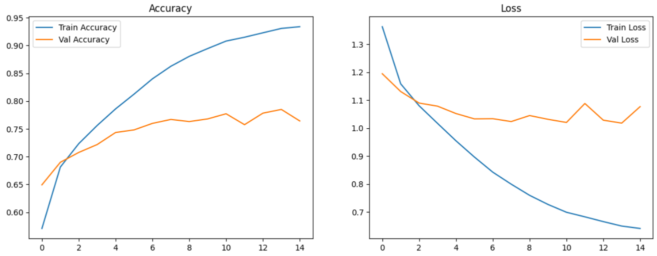

Fig. 2. AffectNet training and validation curves.

### C. Phase 2: FER2013 Fine-tuning

**Dataset:** FER2013 with 7 classes (excluding Contempt). Training: 28,709, Test: 7,178 grayscale 48×48 images [2].

Fig. 3 shows there is a lot of class imbalance as Happy (8,989 samples) is prevalent and Disgust (547 samples) is severely underrepresented, which encourages class weighting during fine-tuning.

**Domain Adaptation:** Grayscale converted to pseudo-RGB via channel triplication.

**Enhanced Augmentation:** RandomRotation (±12), RandomAffine (translate=0.08), ColorJitter (brightness=0.25, contrast=0.25), **RandomErasing** (p=0.15, scale=(0.02, 0.12)).

Fig. 4 illustrates the augmentation pipeline with transformed samples (top row) versus originals (bottom row), including rotation, affine transforms, color jitter, and RandomErasing to enhance robustness.

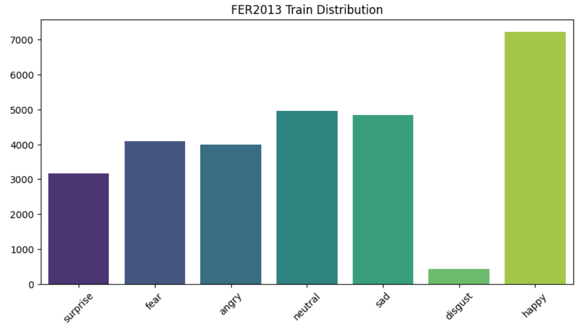

Fig. 3. FER2013 class distribution showing imbalance.

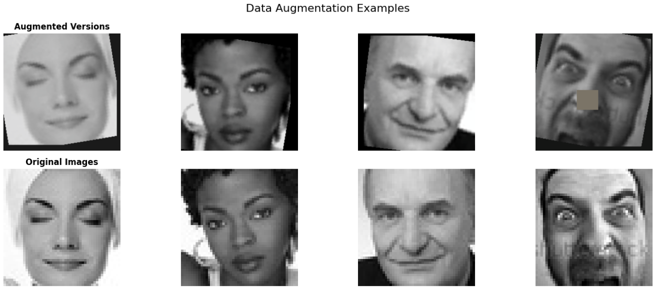

Fig. 4. Augmentation examples: augmented (top) vs. original (bottom).

**Differentiated Learning Rates:** Backbone: $10^{-5}$, Spatial Attention: $5 \times 10^{-5}$, Classifier: $10^{-3}$.

**Training:** 12 epochs with periodic evaluation every 3 epochs, class weighting, label smoothing.

Fig. 5 indicates that domain adaptation is achieved successfully as the validation accuracy increases gradually as the 12 epochs pass and the training loss reduces, which indicates that the transfer learning is successful without any catastrophic forgetting.

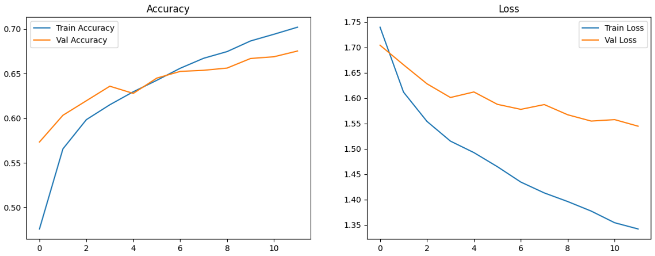

Fig. 5. FER2013 fine-tuning showing domain adaptation.

## IV. Results

### A. FER2013 Evaluation

TABLE I shows detailed per-class performance figures on 7,178 test samples. Happy has the best F1-score (0.87) and the worst F1-score (0.48) with the aggregate accuracy of 68 percent and macro F1-score of 0.65.

TABLE I
Classification Report on FER2013 Test Set

| Emotion | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| Anger | 0.59 | 0.62 | 0.61 | 958 |
| Disgust | 0.51 | 0.73 | 0.60 | 111 |
| Fear | 0.56 | 0.43 | 0.48 | 1024 |
| Happy | 0.89 | 0.85 | 0.87 | 1774 |
| Neutral | 0.62 | 0.68 | 0.65 | 1233 |
| Sad | 0.55 | 0.57 | 0.56 | 1247 |
| Surprise | 0.76 | 0.80 | 0.78 | 831 |
| **Accuracy** | | | **0.68** | **7178** |
| **Macro Avg** | 0.64 | 0.67 | 0.65 | 7178 |
| **Weighted Avg** | 0.68 | 0.68 | 0.67 | 7178 |

### B. Confusion Analysis

The confusion matrix as in Fig. 6 displays the absolute numbers of correct prediction and depicts the high score on the diagonal, but demonstrates the systemic errors, especially Fear-Sad (229 samples) and Sad-Neutral (407 total samples).

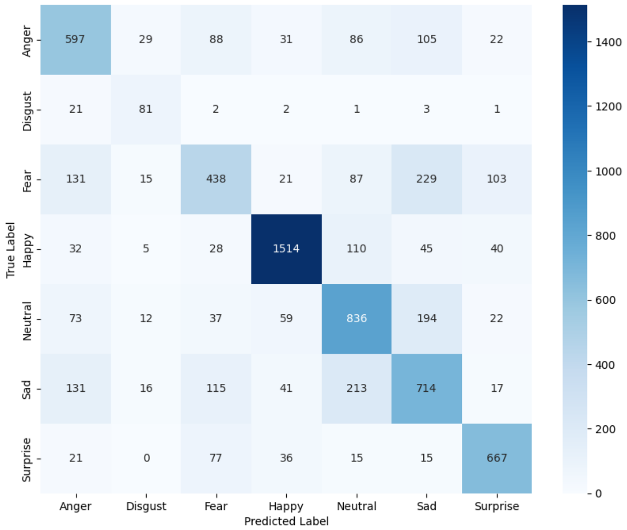

Fig. 6. Confusion matrix on FER2013 test set.

Fig. 7 depicts percentages of confusion with the normalization, where it is evident that the morphological similarity of Fear-Sad is 22.4 which is the highest confusion rate, and this method is more difficult to detect with the low-resolution recognition.

### C. Per-Class Performance

Fig. 8 compares per-class accuracy against the mean baseline (68%). Happy (85%) and Surprise (80%) are far ahead of the average compared to Fear (43%) which is out because there is a high confusion between Sad and Angry.

### D. Precision, Recall, F1-Score

The values of precision, recall, and F1-score of all classes are shown in Fig. 9. Disgust has a high level of recall

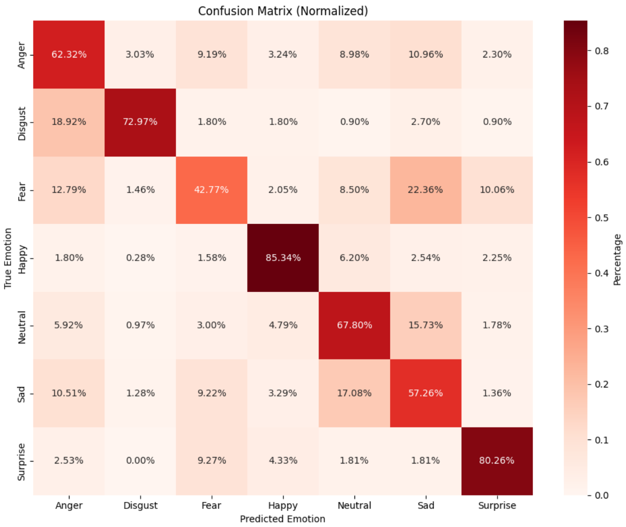

Fig. 7. Misclassification Heatmap.

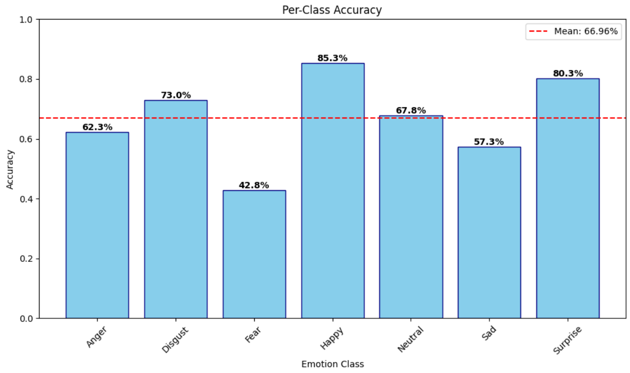

Fig. 8. Per-class accuracy with mean baseline.

(0.73) and low level of precision (0.51) which points to over-predicting and low level of under-detection respectively.

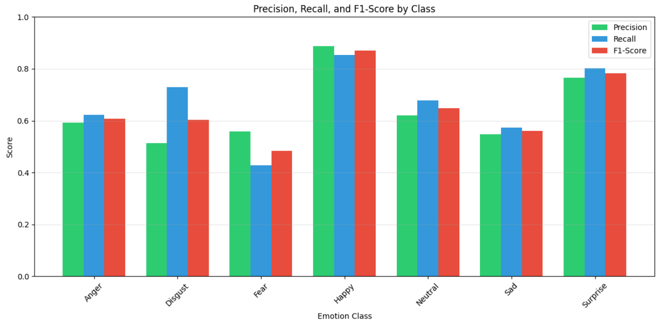

Fig. 9. Precision, Recall, F1-Score comparison.

### E. Statistical Analysis

Table II presents the results of the Chi-square test (χ² = 16557.93, $p \approx 0$), confirming that the patterns of non-random confusion are statistically significant. High-confusion pairs ($>$ 15%) indicate that there are morphologically similar emotions that need specific improvements.

TABLE II
STATISTICAL ANALYSIS OF CONFUSION PATTERNS

| **Statistical Analysis of Confusion Patterns** | |
|---|---|
| Chi-square statistic | 16557.93 |
| P-value | $0.0000 \times 10^0$ |
| Degrees of freedom | 36 |
| **Conclusion** | Confusion patterns are statistically significant ($p < 0.05$) |
| **Analyzing High-Confusion Pairs** | |
| *High Confusion Pairs* ($>$ 15%): | |
| Fear → Sad | 22.4% (229 samples) |
| Disgust → Anger | 18.9% (21 samples) |
| Sad → Neutral | 17.1% (213 samples) |
| Neutral → Sad | 15.7% (194 samples) |
| **Classes Needing Improvement** ($< 50\% accuracy$) | |
| Fear (42.8% accuracy) | Most confused with Sad (22.4%), Anger (12.8%), and Surprise (10.1%) |

Fig. 10 shows some model misclassifications along with their predicted labels, ground truth and confidence values, showing some ambiguity in labels, slight variation in expression that is difficult to see at 48 x 48.

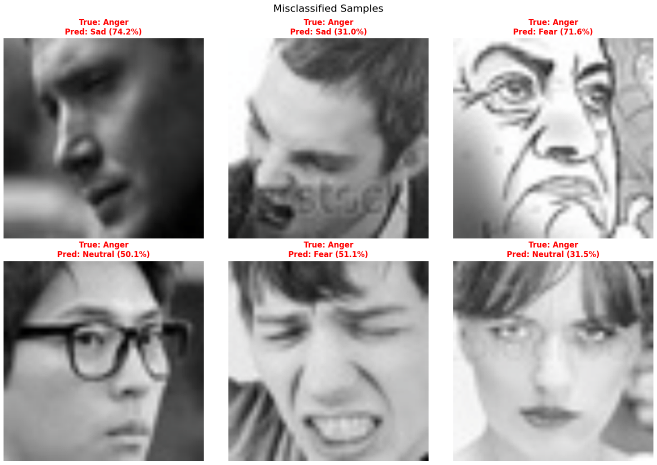

Fig. 10. Misclassified samples with predictions and confidence.

## V. Discussion

### A. Transfer Learning Effectiveness

Two phase training is effective in transferring RGB to grayscale. 41Differentiated learning rates retain such features during higher layer adaptation, and this is given by pre-training on AffectNet.

### B. Attention Mechanism Impact

Channel and spatial attention pay attention to areas of emotion-relevant features (eyes, mouth), enhancing increased recognition of subtle expressions as compared to the baseline ResNet34.

### *C. Confusion Patterns*

**Fear → Sad (22.4%):** Both are depressed mouth and strained muscles. (44) Eye widening is characteristic of Fear, but is a weak signal in the resolution of 48x48. This is the most confused pair that has 229 misclassified samples.

**Disgust → Anger (18.9%):** Share wrapped brows and tensed features. Wrinkle (Disgust) vs. clench of jaw (Anger) at low resolution are hard to tell apart and this leads to 21 misclassifications.

**Sad → Neutral (17.1% and 15.7%):** Controversial with 213 samples confused Sad → Neutral and 194 samples Neutral → Sad, this is due to minor variations in facial relaxation.

**Comparison:** Ma et al. [4]: Our strategy: 68%. accuracy, at the cost of real-time speed of trading, to better identification of subtle expression by deeper feature hierarchies.

### *D. Limitations*

Despite the competitive performance of the proposed framework, a number of limitations remain. The resolution of FER2013 images is 48x48, which limits the acquisition of fine-grained expressions of facial expressions, resulting in a high rate of confusion among morphologically related facial expressions, including Fear-Sad and Sad-Neutral. Moreover, FER2013 has noisy and unclear labels, and this sets a performance limit, especially when dealing with minority classes such as Disgust. Last but not least, ResNet34 with CBAM incurs higher computational cost and does not utilize the dynamic expression of a static image based setup.

### *E. Future Work*

Future work could examine region specific facial branch with late fusion where subtle expressions can be better discriminated. Then, video based FER temporal modeling could be explored with recurrent or transformer based networks to capture expression development. To deal with high confusion emotion pairs, confusion aware training strategies could be implemented, including hard negative mining or focal loss. Moreover, the domain adaptation methods and model compression strategies could be taken into account to enhance cross modal robustness and allow efficient edge deployment.

## VI. Conclusion

This paper suggests a two phase transfer learning method of facial expression recognition, which serves as a valuable and practical way to train models on RGB images and apply them to grayscale environments. With the addition of CBAM based channel and spatial attention modules to ResNet34, layer-wise learning rates and superior data augmentation methods (Random Erasing) the proposed model achieved excellent performance on the FER2013 dataset with a 68 percent classification accuracy and a macro F1-score of 0.65. Statistical analysis pointed out that there has been a consistent confusion between some pairs of emotions, specifically Fear-Sad and Disgust-Anger and this has shown areas where there is a need to improve. This study shows that attention-directed transfer learning has the potential to improve cross domain generalization and is applicable to FER applications in the real world when image modalities are different. Future researchers might concentrate on the concept of temporal modeling of video-based FER, region specific feature projection and model optimization in order to implement edge deployment, further enhancing the ability to detect subtle expressions and efficacy.

## References

[1] Y. Zhang, D. Peng, Y. Wang, and J. Wang, "Research on facial expression recognition algorithm based on deep learning," in *2022 5th World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM)*, 2022.

[2] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio, "Challenges in representation learning: A report on three machine learning contests," in *Neural Information Processing*, M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil, Eds., 2013.

[3] A. Alshammari and M. Alshammari, "Emotional facial expression detection using yolov8," *Engineering, Technology Applied Science Research*, 2024.

[4] Q. Ma, D. Zhang, Z. Shen, H. Zhu, C. Chen, and W. Ma, "Yolo11-ae: An enhanced fine-grained facial expression recognition model with dual-channel attention fusion," in *2025 IEEE International Conference on Pattern Recognition, Machine Vision and Artificial Intelligence (PRMVAI)*, 2025.

[5] A. K. Roy, H. K. Kathania, A. Sharma, A. Dey, and M. S. A. Ansari, "Resemotenet: Bridging accuracy and loss reduction in facial emotion recognition," 2024.

[6] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018.

[7] A. Nadhum, S. Anuar, and S. Ismail, "A comparative study of resnet50 and yolov9 for face detection and gender classification," *Engineering, Technology Applied Science Research*, 2025.

[8] M. M. A. Parambil, L. Ali, M. Swavaf, S. Bouktif, M. Gochoo, H. Aljassmi, and F. Alnajjar, "Navigating the yolo landscape: A comparative study of object detection models for emotion recognition," *IEEE Access*, 2024.

[9] Y. Ma, R. Lu, R. Wanchun, Y. Huang, W. Li, and Y. Wang, "Alf-yolo: a modified yolov8n algorithm for precise emotion detection via facial expressions," *Journal of Real-Time Image Processing*, 2025.

[10] C. Xu, Y. Du, W. Zheng, T. Li, and Z. Yuan, "Facial expression recognition based on yolov8 deep learning in complex scenes," *International Journal of Information and Communication Technology*, 2025.

[11] S. Nathani, "A comparative study of transfer learning for emotion recognition using cnn and modified vgg16 models," 2024.

[12] L. Atymtayeva, M. Kanatov, A. Musleh, and G. Tulemissova, "Fast facial expression recognition system: selection of models," *Appl. Math*, 2023.

[13] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, 2019.