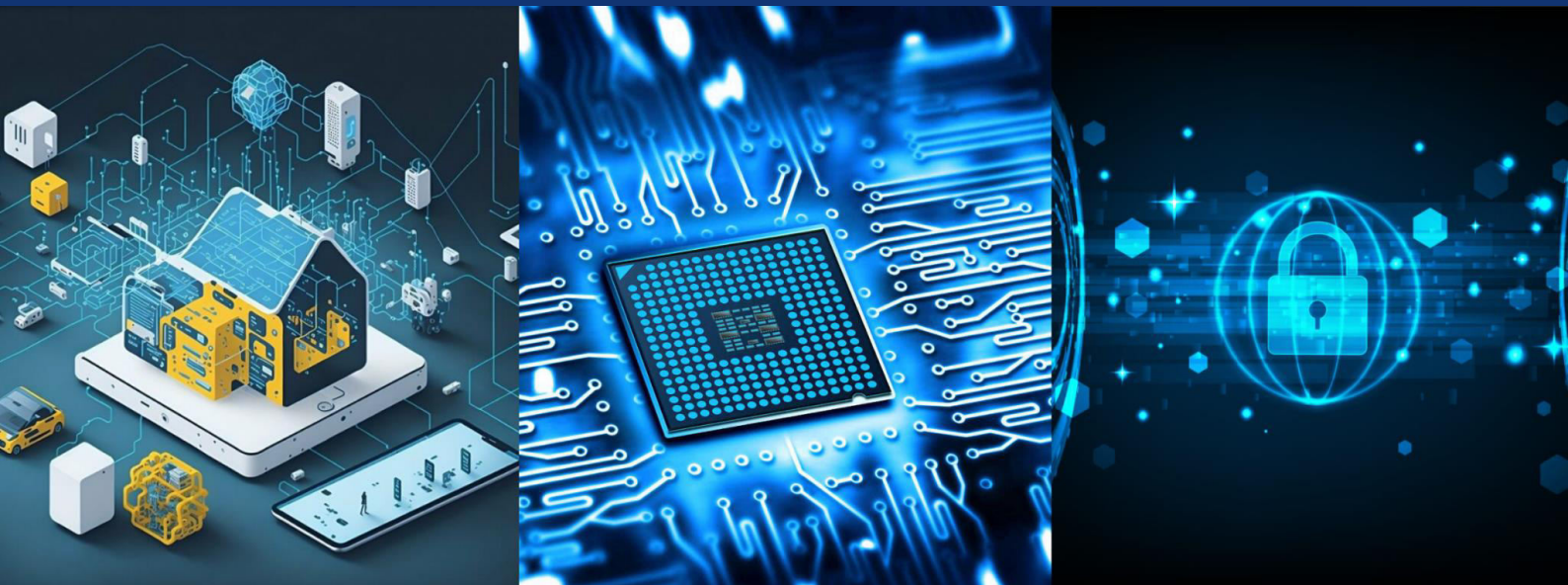
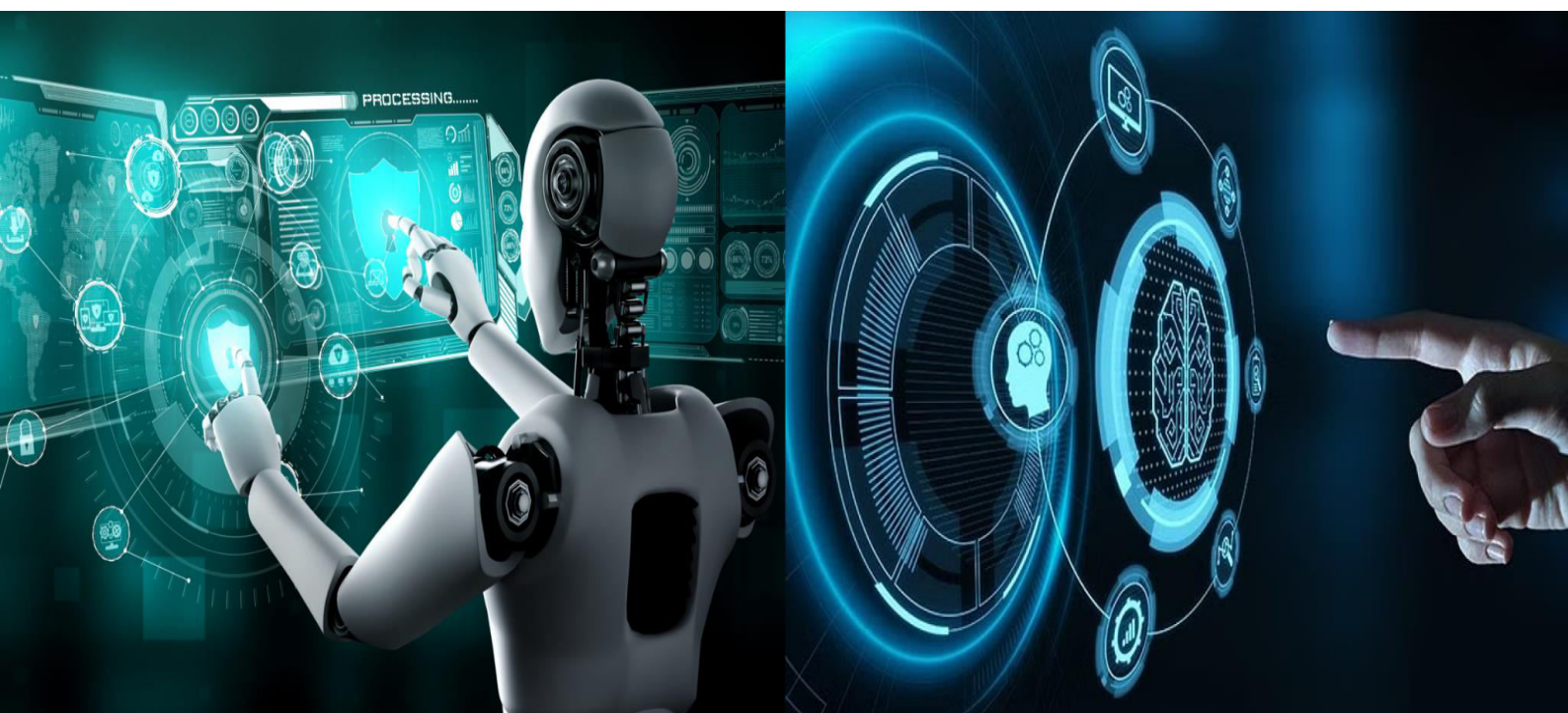




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Confidence-Calibrated Landmark Attention for Lightweight Facial Expression Recognition

Shantha Kumar S¹, Imrana Siddiqua S², Dr. S Mohana³, Dr. V Punitha⁴

Department of Computer Science and Engineering, Saranathan College of Engineering, Panjappur, Tamil Nadu, India

Professor, Department of Computer Science and Engineering, Saranathan College of Engineering, Panjappur,
Tamil Nadu, India

Professor & Head, Department of Computer Science and Engineering, Saranathan College of Engineering, Panjappur,
Tamil Nadu, India

ABSTRACT: Facial Expression Recognition (FER) must classify expressions with limited computational overhead, particularly on resource-constrained devices. Compact CNNs use depthwise-separable convolution and global average pooling, yet occlusion, blur, and lighting changes can make landmark estimates unreliable. This paper proposes Confidence-Calibrated Landmark Attention (CCLA), which combines landmark positions with per-point reliability scores to create a weighted spatial guide for a lightweight CNN. The evaluation plan covers accuracy, Macro-F1, model size, inference latency, and masked-face robustness. Quantitative findings will be added after the proposed experiments are run.

KEYWORDS: facial expression recognition, lightweight CNN, landmark attention, confidence calibration, occlusion robustness

I. INTRODUCTION

FER maps a face image to an expression category and supports applications such as interactive interfaces, driver monitoring, learning environments, and assistive technology. In practical deployment, a model should process images efficiently without losing useful facial cues when lighting, pose, or visibility changes. Zhou et al. describe a compact FER system that combines face localization with residual blocks, depthwise-separable convolution, and global average pooling [1]. A broader review identifies capture variation and training-data limitations as continuing challenges for deep FER [2].

The quality of the facial crop and its geometry affects the evidence available to the classifier. MTCNN estimates face boxes and facial landmarks through a cascaded process [3], while Kazemi and Sullivan present a fast alignment approach based on an ensemble of regression trees [4]. These techniques localize facial structure, but points falling on a covered or visually degraded region may still provide misleading guidance.

Attention can help a network concentrate on informative facial regions rather than treating every spatial location equally. Region Attention Networks address changes in pose and occlusion [5], and an occlusion-adaptive method uses landmark-guided attention to reason about corrupted facial evidence [6]. This work examines a more specific design question: how should the reliability of each detected landmark affect the spatial guidance provided to a lightweight FER model?

We propose Confidence-Calibrated Landmark Attention (CCLA), a module that forms an attention guide from both landmark coordinates and their reliability scores. The scores are normalized so that a less dependable point has a smaller influence on the combined map. The proposed contribution is this point-level reliability weighting within an efficient FER pipeline. Its effect on recognition, robustness, and computational cost remains to be established by experiments.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE SURVEY

Efficiency-oriented FER research seeks to reduce the computation required for expression classification while retaining discriminative facial cues. Zhou et al. report a real-time model with MTCNN, residual modules, depthwise-separable layers, and global average pooling; their study reports 67% FER-2013 accuracy with an 872.9-kB model file [1]. MobileNets similarly use depthwise-separable convolution to limit the cost of mobile vision models [7]. Huo et al. investigate another FER network built around improved depthwise-separable convolution, including Xception-style and inverted-residual components [8]. CCLA keeps the efficient-backbone idea but changes how landmark information contributes to spatial attention.

Landmark localization supplies geometric cues for eyes, nose, and mouth regions. MTCNN predicts face landmarks as part of its detection cascade [3], whereas the regression-tree method of Kazemi and Sullivan provides rapid facial alignment [4]. Position estimates alone are not the same as calibrated, per-point reliability values. CCLA therefore treats reliability estimation as an explicit implementation step; the chosen estimator and its calibration should be documented and validated rather than assumed to be correct.

Attention modules have been used to refine features at different levels. CBAM applies channel attention and spatial attention in sequence [9]. Region Attention Networks weight facial areas to improve tolerance to pose changes and occlusion [5], while Ding et al. describe a network that uses landmark-guided attention for occlusion adaptation [6]. CCLA differs in the construction of its guidance map: it applies normalized reliability weights to individual Gaussian landmark maps before using the result to calibrate CNN features. This is a design distinction to test, not evidence by itself that CCLA performs better.

FER-2013 is a widely used benchmark originating from a machine-learning challenge [10]. FER+ addresses uncertainty in expression labels by preserving crowd-sourced label distributions instead of reducing every image to a single vote [11]. The present evaluation plan uses FER-2013 for the primary comparison and keeps any artificially masked test set separate from the standard test set. Reporting the benchmark split, preprocessing choices, accuracy, Macro-F1, parameter count, and latency will make the comparison easier to interpret.

III. METHODOLOGY / APPROACH

A. Proposed Architecture

Figure 1 summarizes the proposed data path. After face detection and alignment, the resized face is processed by a lightweight CNN to obtain feature map F . In parallel, five facial landmarks and their reliability scores are converted into a spatial guidance map. CCLA joins that map with F , estimates a spatial attention map, and calibrates the features before pooling and seven-class classification.

The backbone uses depthwise-separable convolutions to reduce the cost of spatial filtering and channel mixing, following efficient CNN design principles [1], [7]. The attention module is inserted after the second feature block, where the feature map is $14 \times 14 \times 64$ in the proposed configuration.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

CCLA-FER: proposed end-to-end flow

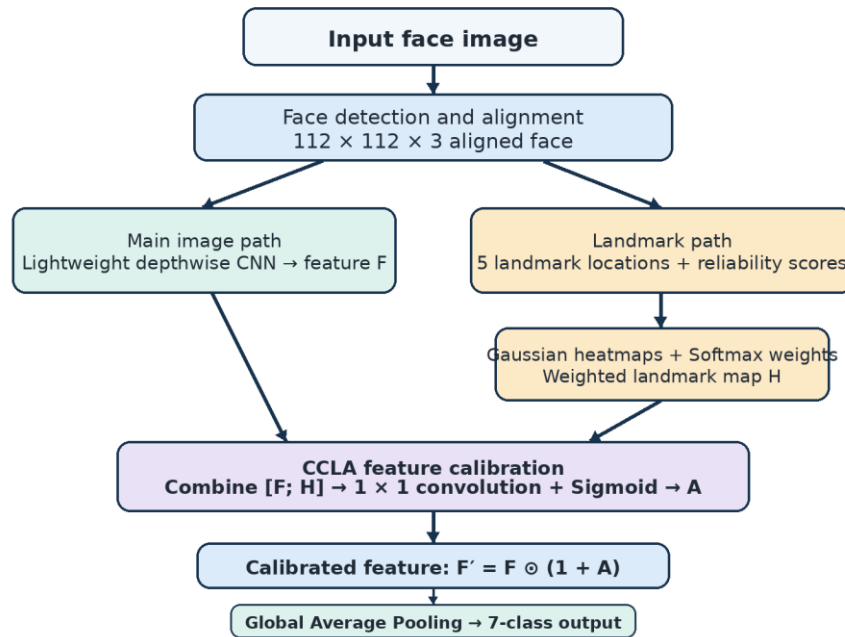


Fig.1. Proposed end-to-end CCLA-FER pipeline. The diagram is a conceptual architecture, not a measured result.

B. Landmark and Reliability Extraction

The design uses five points: the two eye locations, the nose, and the two mouth corners. Let $L = \{l_1, l_2, l_3, l_4, l_5\}$ denote the coordinates and $R = \{r_1, r_2, r_3, r_4, r_5\}$ their reliability scores, with $0 \leq r_i \leq 1$. The scores may be supplied by the landmark detector. If only landmark heatmaps are available, a normalized peak response can be used as a provisional proxy, but it should not be treated as a calibrated probability without validation. Figure 2 is a schematic of the five-point representation and the intended effect of reliability weighting.

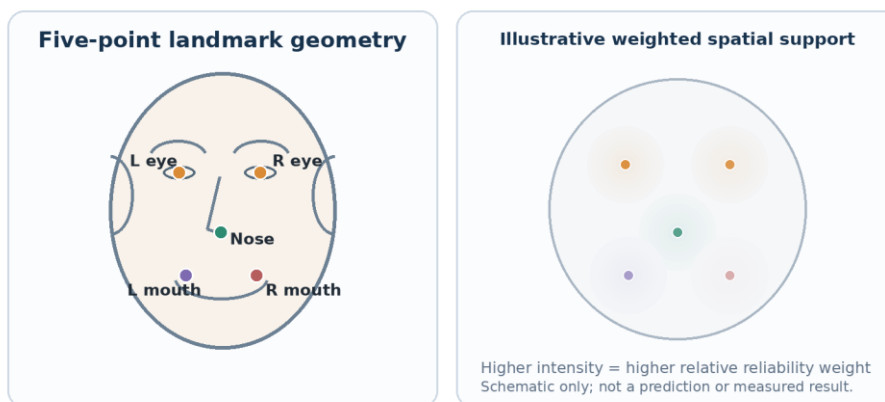


Fig.2. Five-point facial landmark representation and illustrative reliability-weighted spatial support (schematic only)

C. Confidence-Calibrated Landmark Attention

For each landmark $l_i = (x_i, y_i)$, a Gaussian map marks its local neighborhood:

$$H_i(x, y) = \exp(-((x - x_i)^2 + (y - y_i)^2) / (2\sigma^2))$$



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The reliability scores are converted into normalized weights using a temperature-scaled Softmax:

$$w_i = \exp(r_i / \tau) / \sum_j \exp(r_j / \tau)$$

The temperature parameter τ determines how strongly score differences affect the normalized weights. The combined spatial guide is defined as:

$$H(x, y) = \sum_i w_i H_i(x, y)$$

A point with lower reliability still contributes to the map, but its share is reduced relative to points assigned higher weights.

D. Feature Calibration

Let F be the feature map produced by the compact CNN. The map H is resized to match F and concatenated with the image features. A 1×1 convolution and Sigmoid activation then estimate a spatial attention map:

$$A = \sigma(\text{Conv}_{1 \times 1}([F; H]))$$

The calibrated feature map is:

$$F' = F \odot (1 + A)$$

The residual-style modulation scales the original feature response according to A , rather than replacing F with the attention map alone. Figure 3 summarizes these operations.

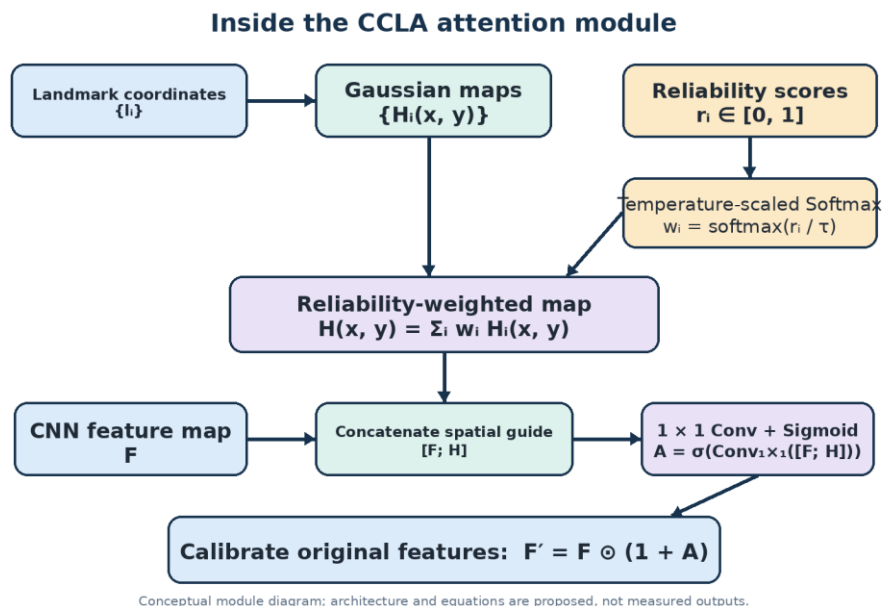


Fig.3. Internal operations of the proposed Confidence-Calibrated Landmark Attention module

E. Global Average Pooling and Classification

Global Average Pooling converts the calibrated feature map into a feature vector. A linear layer then predicts one of seven FER-2013 categories: anger, disgust, fear, happiness, sadness, surprise, or neutral.

F. Dataset and Preprocessing

FER-2013 is the primary dataset for this study [10]. For a separate robustness evaluation, a fixed eye or mouth region can be masked in selected test images. The masked set is an additional stress test and should not be merged with the



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

standard test results. FER+ is relevant background because its crowd-sourced annotations illustrate that expression labels can be ambiguous [11].

Before training, the face is detected, landmarks are estimated, the crop is aligned, and the image is resized to 112×112 pixels. Augmentation is planned to include horizontal flips, small rotations, brightness changes, and random occlusion. The same preprocessing policy should be applied consistently to each model in the ablation.

G. Network Structure

TABLE.I. NETWORK STRUCTURE

Stage	Output	Operation
Input	$112 \times 112 \times 3$	RGB face
Stem	$56 \times 56 \times 32$	3×3 convolution
Block 1	$28 \times 28 \times 64$	Depthwise-separable convolution
Block 2	$14 \times 14 \times 64$	Depthwise-separable convolution
CCLA	$14 \times 14 \times 64$	Reliability-aware landmark attention
GAP	64	Global Average Pooling
Classifier	7	Linear layer

H. Training Configuration

TABLE.II. TRAINING CONFIGURATION

Parameter	Setting
Framework	PyTorch
Input size	112×112
Optimizer	AdamW
Loss	Cross-Entropy
Batch size	64
Weight decay	1×10^{-4}
Temperature τ	Selected using validation data

IV. RESULTS & DISCUSSION

No completed training run or measured evaluation result is available in this manuscript version. This section therefore records the planned experiments and the criteria that will be used to assess CCLA; numerical performance claims should be inserted only after the experiments have been performed.

A. Ablation Study

Three versions will be compared so the additional value of reliability weighting can be separated from the effect of ordinary landmark guidance:



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE.III. ABLATION STUDY

Model	Description	Purpose
Baseline CNN	Lightweight depthwise CNN	Measure baseline performance
Landmark Attention	Uses landmark locations	Measure ordinary landmark guidance
CCLA	Uses landmark locations and reliability	Measure confidence calibration

B. Evaluation Metrics

Recognition quality will be measured by accuracy and Macro-F1, with the latter giving each expression class equal weight in the final score. Parameter count and inference latency will be reported alongside these metrics to show whether the proposed attention module remains suitable for lightweight use.

C. Robustness Evaluation

The standard FER-2013 test split should be evaluated once without synthetic masking and once with fixed eye-region or mouth-region masks. Comparisons between Landmark Attention and CCLA on masked images will test whether per-point confidence weighting helps when some facial evidence is obscured. Results for clean and masked images should be shown separately.

D. Discussion

The design hypothesis is that reliability-aware weighting may reduce the influence of unreliable landmarks under occlusion. Evidence for this claim would require a measured improvement over ordinary landmark attention on the masked set, together with a small increase in parameter count and latency. Until those experiments are completed, this remains a hypothesis rather than a finding.

V. CONCLUSION

This paper introduces Confidence-Calibrated Landmark Attention (CCLA), a proposed reliability-aware spatial guidance module for lightweight facial expression recognition. Instead of combining landmark maps with equal importance, it normalizes per-point reliability scores and uses them to form a weighted map that interacts with CNN features.

The proposed pipeline retains an efficient depthwise-separable backbone and global average pooling, while placing CCLA between feature extraction and classification. A planned comparison of the baseline CNN, conventional landmark attention, and CCLA under clean and masked conditions will determine whether the new weighting strategy improves robustness without an unacceptable increase in model size or inference time. Those conclusions must be based on completed experiments.

REFERENCES

- [1] N. Zhou, R. Liang, and W. Shi, "A Lightweight Convolutional Neural Network for Real-Time Facial Expression Detection," IEEE Access, vol. 9, pp. 5573–5584, 2021, doi: 10.1109/ACCESS.2020.3046715.
- [2] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, Jul.–Sep. 2022, doi: 10.1109/TAFFC.2020.2981446.
- [3] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, Oct. 2016, doi: 10.1109/LSP.2016.2603342.
- [4] V. Kazemi and J. Sullivan, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1867–1874, doi: 10.1109/CVPR.2014.241.
- [5] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," IEEE Transactions on Image Processing, vol. 29, pp. 4057–4069, 2020, doi: 10.1109/TIP.2019.2956143.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

- [6] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-Adaptive Deep Network for Robust Facial Expression Recognition," in Proceedings of the 2020 IEEE/IAPR International Joint Conference on Biometrics (IJCB), 2020, pp. 1–9, doi: 10.1109/IJCB48548.2020.9304923.
- [7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," arXiv:1704.04861, 2017.
- [8] H. Huo, Y. Yu, and Z. Liu, "Facial Expression Recognition Based on Improved Depthwise Separable Convolutional Network," Multimedia Tools and Applications, vol. 82, pp. 18635–18652, 2023, doi: 10.1007/s11042-022-14066-6.
- [9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.
- [10] I. J. Goodfellow et al., "Challenges in Representation Learning: A Report on Three Machine Learning Contests," Neural Networks, vol. 64, pp. 59–63, 2015, doi: 10.1016/j.neunet.2014.09.005.
- [11] E. Barsoum, C. Zhang, C. Canton Ferrer, and Z. Zhang, "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 279–283, doi: 10.1145/2993148.2993165.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details