DYNAMIC FACIAL EXPRESSION RECOGNITION IN THE WILD USING MAMBA-STYLE SELECTIVE SSM AND FACIAL ATTENTION MECHANISM

Yudhistira Arditya Pratama, Theophilus Ezra Nugroho Pandin, Yi-Zeng Hsieh

Department of Electrical Engineering,
National Taiwan University of Science and Technology, Taiwan
*E-mail: yzhsieh@mail.ntust.edu.tw

ABSTRACT

Dynamic facial expression recognition (FER) in the wild remains challenging due to factors such as occlusions, motion blur, head pose variations, and highly imbalanced emotion classes (e.g., "Disgust" and "Fear" are notably rare in the DFEW dataset). In this work, we propose an efficient FER framework that integrates a lightweight CNN front end with Mamba-style Selective State Space Model (SSM) blocks for linear-time temporal modeling. This architecture is further enhanced by a facial CBAM module—incorporating channel, spatial, and region-aware attention mechanisms—and a contrast-enhancement preprocessing step. When trained on the DFEW dataset using class-balanced focal loss and the OneCycleLR schedule, the proposed method achieves a validation weighted average recall (WAR) of 66% and an unweighted average recall (UAR) of 54%. Grad-CAM visualizations verify that the model accurately focuses on critical facial regions such as the eyes, mouth, and cheeks under diverse conditions. By effectively capturing long-range temporal dynamics and subtle facial cues without incurring quadratic computational costs, the proposed approach demonstrates strong potential for real-world applications. Future research will explore advanced class-balancing techniques and adaptive SSM tuning to further improve recognition performance, particularly for underrepresented emotions.

Keywords: *Dynamic facial expression recognition, Mamba, Selective SSM, Facial Attention Mechanism, Deep Learning.*

1. INTRODUCTION

Several recent works have highlighted the challenges of extracting rich facial features from videos under real-world conditions. Li et al [1], note that driver emotion datasets often suffer from uneven class distributions and neglect temporal context, while Ma et al. [2] demonstrate that standard CNN-based dynamic FER methods ignore long-range dependencies and struggle with in-the-wild noise. Chen et al. [3] show the benefit of leveraging landmark-aware image models but require full fine-tuning for video data, and Li et al. [4] integrate convolutional and state-space modules for efficient long-sequence analysis yet still rely on careful hand-tuning. Beyond FER, Partha [5] applies Mamba for micro-gesture recognition to handle longer videos more efficiently, and Ma et al. [6] fuse YOLO detection with selective state-space attention to focus on key facial regions. Together, these studies motivate our design: a lightweight CNN front end to capture local cues, followed by Mamba-style blocks for linear-cost, long-range temporal modeling, plus region-focused attention and class-weighted losses to counteract DFEW's imbalance.

2. RELATED WORKS

Early methods for dynamic facial expression recognition (FER) mainly relied on sequential models combining CNNs and RNNs. In these approaches, CNNs extracted spatial features from each frame while recurrent units such as LSTMs modeled temporal changes across sequences. Although this pipeline worked in controlled conditions, it separated spatial and temporal cues, leading to information loss when expressions evolved slowly or when faces were blurred [7]. To overcome this limitation, researchers later extended CNNs into 3D, enabling joint spatio-temporal learning. While 3D CNNs captured motion patterns more effectively, most of them remained shallow and struggled with subtle expressions in real-world, in-the-wild settings [8].

The introduction of Transformers marked a new stage in FER research. By applying self-attention, these models could capture global spatio-temporal dependencies in a single step, which improved recognition accuracy compared to CNN–RNN or 3D CNN architectures. In particular, spatio-temporal Transformer models achieved strong performance on benchmarks such as DFEW, showing the benefits of long-range context modeling. However, their quadratic memory and computational costs limited their practicality for deployment in real-time or resource-constrained environments [2], [3].

More recently, state space models (SSMs) have emerged as an efficient alternative for temporal modeling. Unlike Transformers, SSMs update hidden states in parallel for each timestep, achieving linear complexity while still modeling long-range dynamics. Variants such as Face Mamba and FER-YOLO-Mamba have shown that combining SSM blocks with lightweight

979-8-3315-7206-8/25/$31.00

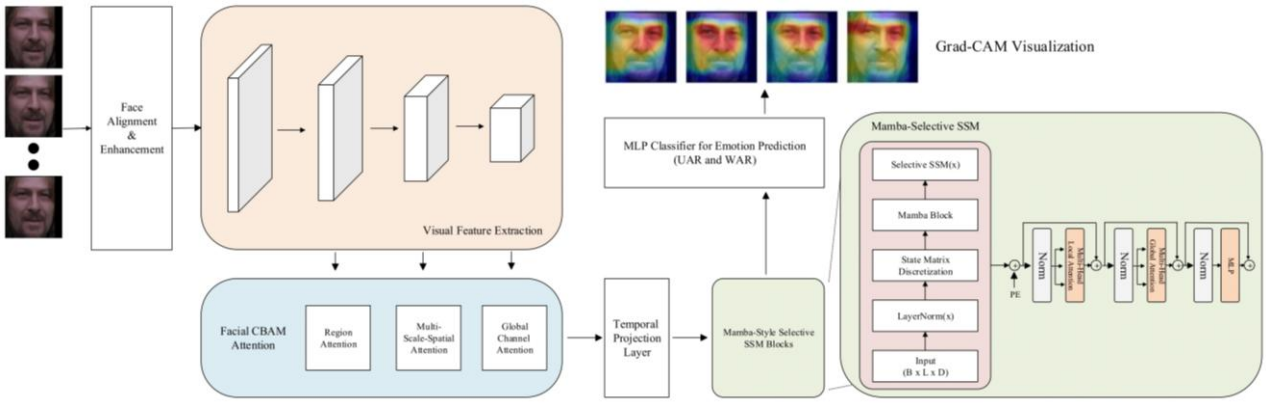

Fig. 1 The overall architecture of the proposed dynamic facial expression recognition framework

attention can focus on key facial regions and improve recognition of subtle or rare emotions, including disgust and fear [4], [5], [6]. These results highlight the potential of SSMs as a faster yet effective solution for dynamic FER.

In parallel, another research direction has focused on how to represent facial information effectively for emotion recognition. Some studies used entire facial frames as model input, but this led to high computational cost, especially in real-time scenarios [7]. Others restricted the input to local facial regions—such as the eyes, nose, and mouth—reducing complexity but losing global context [9]. To balance efficiency and completeness, later works explored global keypoint representations, which preserved structural information while remaining compact. However, these methods often overlooked fine texture cues necessary for subtle emotion recognition [1], [8].

3. METHODOLOGY

We handle DFEW's severe class imbalance by tallying all training clips into its seven emotions and computing smoothed inverse-frequency weights (clipped to [0.1,10]) so that Disgust, Fear, and Surprise receive higher loss weight. Each fold's CSV lists video names and labels; we auto-detect the correct clip directory via common zero-padding and file-extension patterns. From every video, we uniformly sample 16 frames (repeating or padding when too short), detect and align the largest face with Haar cascades plus eye-based tilt correction, then enhance it via CLAHE, bilateral filtering, and a light unsharp mask. During training, frames undergo mild flips, affine jitter, brightness/contrast tweaks, CLAHE/noise/blur, and color jitter; validation/test use only CLAHE and ImageNet normalization.

As we seen in table 1 and figure 1, our network uses EfficientNet-V2-S (feature extractor only), followed by a three-stage face-tailored CBAM: global channel attention; facial-region attention over eight learned maps (eyes, mouth, etc.) with a landmark mask; and multi-scale spatial attention via 3×3, 5×5, and 7×7 convolutions. Frame features are pooled to 1,280 D, projected to model dimension, and pass through two SelectiveSSM ("Mamba") blocks—each combining depthwise conv with a linear-time state-space scan. A temporal multi-head-self-attention refines inter-frame cues before averaging over 16 frames and classifying via an MLP (1280→512→256→128→7).

We train up to 200 epochs (batch 8) with AdamW (LR 1e-4, decay 0.01), OneCycleLR, and FocalLoss (γ=2) weighted by class. An EMA of weights stabilizes validation, and we early-stop after 20 epochs without UAR improvement. The best EMA snapshot is evaluated on validat1ion, reporting WAR/UAR and per-class metrics, and Grad-CAM confirms attention to key expressive regions.

Table 1. Network architecture for dynamic facial recognition framework using mamba-style selective SSM

No	Module	Shape
1.	Input video	(B, T, 3, H, W)
2.	EfficientNetV2-S backbone	(B·T, 3, H, W) → (B·T, 1280, H`, W`)
3.	CBAM + landmark attention + facial enhancer	(B·T, 1280, H`, W`) → (B·T, 1280, H`, W`)
4.	Global average pooling	(B·T, 1280, H`, W`) → (B·T, 1280)
5.	Reshape to sequence	(B·T, 1280) → (B, T, 1280)
6.	Temporal Projection (linear)	(B, T, 1280) → (B, T, 1280)
7.	Mamba (SSM) blocks	(B, T, 1280) → (B, T, 1280)
8.	Multi-head temporal attention	(B, T, 1280) → (B, T, 1280)
9.	Temporal pooling (Avg1D)	(B, T, 1280) → (B, 1280)
10.	Classifier MLP	1280 → 512 → 256 → 128 → 7

The proposed Landmark-Enhanced EfficientFER with SSM model takes video inputs of size *(B, T, 3, H, W)*. Each frame is processed by a pretrained EfficientNetV2-S backbone, producing spatial features of (B · T, 1280, H`, W`). These are refined through a landmark-enhanced CBAM, a landmark-constrained attention block, and a facial enhancement layer that emphasize discriminative facial regions. A global average pooling step compresses spatial dimensions, yielding per-frame descriptors of size 1280, which are then reshaped into a temporal sequence (B, T,1280).

Table 2. Comparison with state-of-the-art on DFEW dataset

Method	Accuracy of Each Expressions (%)							Metrics (%)		FLOPs (G)
	Happy	Sad	Neutral	Angry	Surprise	Disgust	Fear	UAR	WAR	
R(2+1)D18 [10]	79.67	39.07	57.66	50.39	48.26	3.45	21.06	42.79	53.22	42.36
3D ResNet18 [11]	73.13	48.26	50.51	64.75	50.10	0.00	26.39	44.73	54.98	8.32
ResNet18+LSTM [12]	78.00	40.65	53.77	56.83	45.00	**4.14**	21.62	42.86	43.08	**7.78**
ResNet18+GRU [12], [13]	82.87	63.83	65.06	68.51	52.00	0.86	30.14	51.68	64.02	**7.78**
FormerDFER [14]	84.05	62.57	67.52	70.03	56.43	3.45	**31.78**	53.69	65.70	9.11
R(2+1)D ResNet18 Hybrid [15]	81.27	55.07	58.90	60.66	51.53	3.45	22.88	48.35	59.28	47
Mamba-Style SSM (Ours)	**84.90**	**65.70**	62.90	**70.60**	**54.10**	0.00	24.30	**54.28**	**66.81**	15.2

For temporal modeling, the sequence is projected through a linear layer and processed by stacked Mamba (Selective State Space Model) blocks to capture long-range dependencies. Multi-head temporal attention further enhances inter-frame relations, followed by adaptive temporal pooling to obtain a compact video-level feature vector (B,1280). The classifier MLP (1280 $\rightarrow$ 512 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 7) produces emotion logits for seven classes. In parallel, a Comprehensive *Action Unit* (AU) Analyzer leverages 68-point landmarks to compute frame-level AU scores for interpretability and re-weighting, without altering the main feature flow.

Computational cost. The Mamba-style SSM runs in linear time over frames, while the facial CBAM stack operates per frame only on spatial maps. Concretely, after EfficientNetV2-S produces $(B \cdot T, C, H', W')$, the channel/region/landmark-guided spatial attentions use a few 1×1 and smarll-kernel $k \in \{3,5,7\}$ convolutions on the pooled 2-channel input (avg/max), giving a per-clip cost $O\left(BT(CH'W' + k^2H'W')\right)$ with no quadratic dependence on sequence length. The Mamba block then processes the temporally projected tokens $x_t \in R^D$ using a depthwise 1-D conv $O(B\ T\ D\ d_{conv})$ followed by the selective scan, are given by:

$$\tilde{A} = exp(\Delta t \odot \mathrm{A}),$$
$$h_t = \tilde{A} \odot h_{t-1} + \tilde{B}_t \odot x_t,$$
$$y_t = {C_t}^T h_t + D \odot x_t$$

where $A \in R^N$ (state size N = d_{state}) is diagonalized per channel and $\tilde{B}_t$, C_t are frame-dependent gates obtained from linear projections. This recurrence costs *O (B T D N)*, i.e., linear in *T*. The final temporal MHSA is applied once on $T = 16$ tokens, costing $O(B\ H\ T^2\ d_h)$ and is modest because it runs after heavy spatial pooling. In summary, we compute facial attention only in the spatial dimension (per frame) while Mamba captures temporal dynamics (with light channel wise mixing); by avoiding joint spatiotemporal attention—whose complexity would scale like $O((TH'W')^2)$—the method achieves a significant reduction in compute and memory versus simultaneously attending over both space and time.

4. EXPERIMENT AND RESULTS

We trained the proposed selective-SSM-EfficientFER on the DFEW Fold 1 dataset and applied early stopping at 60 epochs. As illustrated in Fig. 2, both training and validation losses decreased steadily to around 0.5 and 1.0, while the Dice loss dropped to approximately 0.1 and 0.5. Validation WAR increased rapidly, reaching 60% within the first 15 epochs, and UAR rose to 54%. Performance continued to improve gradually, achieving 66.81% WAR by epoch 40, with Macro F1 peaking at 53% at epoch 40. Class-wise results showed accuracies of 84.90% for Happy, 65.70% for Sad, 62.90% for Neutral, 70.60% for Angry, 54.10% for Surprise, but only 0.00% for Disgust and 24.30% for Fear.

Visualization. These outcomes highlight the severe under-representation of Disgust and Fear, as well as the frequent confusion between rare categories and more common neutral–angry expressions, as presented in Fig. 5. Furthermore, the Grad-CAM visualizations in Fig. 2 reveal that the backbone generally attends to the upper face, while the facial-CBAM module enhances focus on the cheeks, mouth, and eyes. The refinement module further sharpens attention across frames, enabling the model to capture subtle cues such as tears and downturned lips in challenging sequences.

In the figure 3 and figure 4 show the Grad-CAM temporal evolution for a correctly predicted Happy clip. Across sampled frames, the heatmaps consistently highlight the upper- and mid-face regions, especially the eyes and mouth corners, which are canonical action units for smiles (AU6 + AU12). Early frames (e.g., Frame 3) show localized peaks around the eye muscles as the smile begins to form, while later frames (Frames 11–16) exhibit broader activation covering both cheeks and the mouth, capturing the sustained lip-corner puller. This progression demonstrates that the selective-SSM preserves temporal consistency: the CBAM attends to per-frame discriminative cues, while the Mamba blocks ensure smooth propagation across time, yielding coherent attention to smile dynamics rather than frame-by-frame noise. These visualizations confirm that the model focuses on physiologically meaningful cues for

Happy and that the temporal backbone refines attention as the expression evolves.

While the training and validation curves (Fig. 2) show stable convergence with validation WAR saturating at ~66% and UAR ~54%, class-level breakdown reveals strong imbalance effects. The confusion matrix (Fig. 5) highlights that Disgust and Fear remain the weakest classes. For Disgust, most clips are confused with Neutral or Surprise—consistent with subtle expressions (AU9 nose wrinkler, AU15/16 lip depressors) that are often subdued in low-resolution video. The model tends to misinterpret them as the absence of strong emotion (Neutral) or as Surprise when the mouth opening dominates.

Error Analysis. Fear is even more problematic: over half of Fear clips are misclassified as Surprise or Sad. This is physiologically understandable since Fear and Surprise share overlapping action units (AU1/2/5/26 for brow raise, lid raise, and jaw drop), making them difficult to separate without reliable temporal cues. Moreover, Fear clips are underrepresented in DFEW, so the smoothed inverse-frequency weighting only partially compensates. The validation curves (lower plateau in UAR vs. WAR) confirm that minority classes drag down unweighted averages despite good performance on dominant classes like Happy and Sad.

Together, the training dynamics and confusion patterns emphasize the dataset's skew: the model can learn robust, separable embeddings for frequent classes, but rare, visually overlapping classes like Disgust and Fear remain prone to misclassification. This suggests that augmenting with AU-guided reweighting or temporal contrastive learning could further disambiguate these challenging expressions.

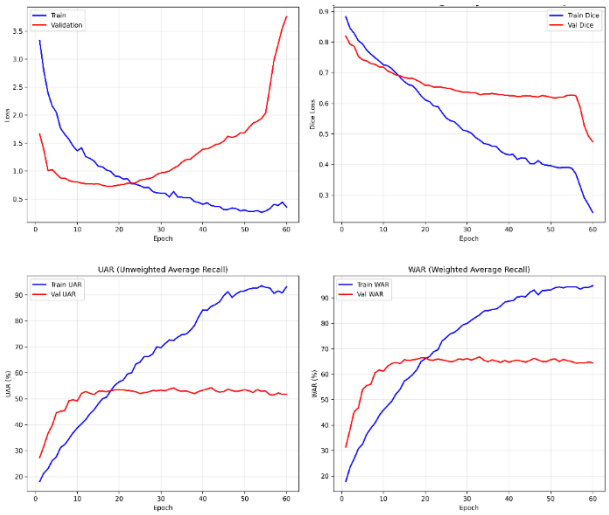

Fig. 2. DFEW Fold 1 Facial-Feature Training Progress

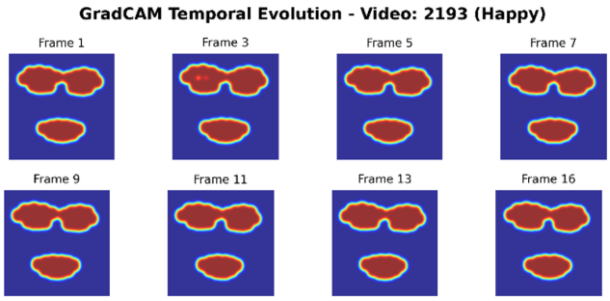

Fig. 3. Visualization Grad-Cam Temporal evolution in each frame

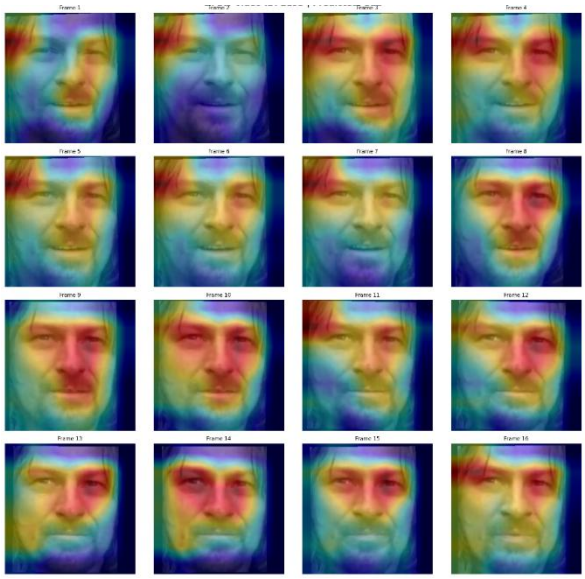

Fig. 4. Sequential Grad-CAM visualization.

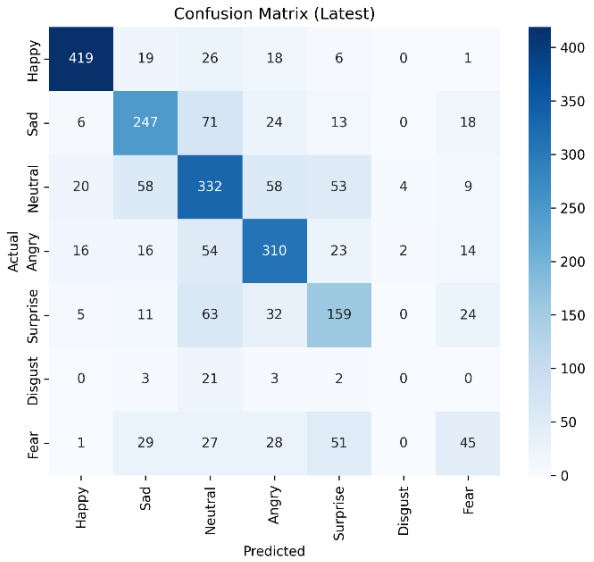

Fig. 5. DFEW Fold 1 Confusion Matrix

5. CONCLUSION

We have shown that integrating lightweight spatio-temporal SSM blocks with facial-CBAM and enhancement layers yields a practical FER model: it converges reliably, captures long-range dynamics, and

localizes fine facial cues. Training on DFEW demonstrated stable optimization, with validation WAR and UAR improving steadily and reaching 66.8% and 54% respectively, while Macro-F1 peaked at 53%. Class-wise analysis revealed strong results for frequent emotions such as Happy, Sad, Neutral, and Angry, but very limited accuracy for Disgust and Fear, reflecting both class imbalance and visual overlap with related expressions. The confusion matrix confirmed that Disgust is often misclassified as Neutral or Surprise, and Fear is frequently confused with Surprise or Sad due to shared action units such as brow raises and jaw drops. Grad-CAM visualizations further verified that the backbone attends broadly to the upper face, while the landmark-enhanced CBAM consistently emphasizes the eyes, cheeks, and mouth. Over time, the selective-SSM preserved temporal consistency, refining attention to evolving cues such as lip pulling or downturned corners.

Although the model achieves reliable convergence and competitive performance on majority classes, the analysis highlights key limitations in handling rare categories. Future work will explore stronger class-balancing strategies, larger and more diverse in-the-wild corpora, and adaptive SSM configurations to improve recognition of low-frequency expressions. Incorporating AU-guided reweighting and temporal contrastive learning may also help disambiguate subtle differences between visually overlapping emotions such as Fear and Surprise. Together, these directions hold promise for building more robust and fair FER systems applicable to unconstrained real-world settings.

REFERENCES

[1] J.Li, L.Yang, C.Lv, Y.Chu, andY.Liu, "GLF-STAF: A Global-Local-Facial Spatio-Temporal Attention Fusion Approach for Driver Emotion Recognition," *IEEE Trans. Consum. Electron.*, 2025.

[2] F.Ma, B.Sun, andS.Li, "Spatio-temporal transformer for dynamic facial expression recognition in the wild," *arXiv Prepr. arXiv2205.04749*, 2022.

[3] Y.Chen, J.Li, S.Shan, M.Wang, andR.Hong, "From Static to Dynamic: Adapting Landmark-Aware Image Models for Facial Expression Recognition in Videos," *IEEE Trans. Affect. Comput.*, vol. PP, pp. 1–15, 2024, doi: 10.1109/TAFFC.2024.3453443.

[4] L.Li, Q.Sun, L.Zhao, H.Sun, F.Zhao, andB.Gu, "Face Mamba: A Facial Emotion Analysis Network Based on VMamba," in *2024 7th International Conference on Machine Learning and Natural Language Processing (MLNLP)*, IEEE, 2024, pp. 1–5.

[5] D. H.Partha, "Micro-gesture recognition using Mamba," 2025.

[6] H.Ma, S.Lei, T.Celik, andH.-C.Li, "Fer-yolo-mamba: Facial expression detection and classification based on selective state space," *arXiv Prepr. arXiv2405.01828*, 2024.

[7] L.Yang, Y.Tian, Y.Song, N.Yang, K.Ma, andL.Xie, "A novel feature separation model exchange-GAN for facial expression recognition," *Knowledge-Based Syst.*, vol. 204, p. 106217, 2020.

[8] W.Shulei *et al.*, "Road rage detection algorithm based on fatigue driving and facial feature point location," *Neural Comput. Appl.*, vol. 34, no. 15, pp. 12361–12371, 2022.

[9] W.Li *et al.*, "Review and perspectives on human emotion for connected automated vehicles," *Automot. Innov.*, vol. 7, no. 1, pp. 4–44, 2024.

[10] D.Tran, H.Wang, L.Torresani, J.Ray, Y.LeCun, andM.Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[11] K.Hara, H.Kataoka, andY.Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[12] K.He, X.Zhang, S.Ren, andJ.Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] J.Chung, C.Gulcehre, K.Cho, andY.Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv Prepr. arXiv1412.3555*, 2014.

[14] Z.Zhao andQ.Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 1553–1561.

[15] G.Sathisha, C. K.Subbaraya, andG. K.Ravikumar, "Facial Expression Recognition in Video Using 3D-CNN Deep Features Discrimination," in *2024 3rd International Conference for Innovation in Technology (INOCON)*, IEEE, 2024, pp. 1–6.