LightMamba: A Multimodal Audio-Visual Framework for Continuous Emotion Recognition

Yuheng Liang[1], Feng Liu[2,†], Yu Yao[3], Mingzhou Liu[3], Jing Yuan[3]
School of Communications and Information Engineering
Jiangsu Key Laboratory of Intelligent Information Processing and Communication Technology
Nanjing University of Posts and Telecommunications, Nanjing, China
{1023010418, 1023010419, 1023162807, 1223014230}@njupt.edu.cn
[†]Corresponding author: liuf@njupt.edu.cn

***Abstract*—Continuous emotion recognition (CER) in the valence-arousal (VA) space is a central task in affective computing and human-computer interaction. While recent deep learning approaches have achieved notable progress using visual inputs, unimodal systems often fail to capture the full spectrum of human emotions. In this paper, we propose LightMamba, a novel multimodal audio-visual framework designed for efficient and accurate CER. Our method integrates visual features extracted from a pretrained Masked Autoencoder (MAE) with audio representations from both VGGish and openSMILE. These fused features are processed by a hybrid temporal modeling architecture comprising a Temporal Convolutional Network (TCN) and a lightweight Mamba block for long-range sequence learning. Experiments on the Aff-Wild2 dataset demonstrate that LightMamba achieves state-of-the-art performance with an average CCC of 0.637, outperforming existing methods while maintaining lower GPU memory usage and competitive training speed. These results highlight the potential of Mamba-based architectures for scalable and deployable affective computing systems.**

***Index Terms*—Continuous Emotion Recognition, Mamba, Valence-Arousal, Audio-Visual Learning**

I. Introduction

Affective computing has emerged as a key area in human-computer interaction (HCI), aiming to endow machines with the ability to perceive, understand, and respond to human emotions [1]. Among various emotion recognition paradigms, continuous emotion recognition (CER) in the Valence-Arousal (VA) space has gained increasing attention, offering a richer and more flexible emotional representation than traditional discrete classifications [2]. In this framework, valence measures the positivity or negativity of an emotion, while arousal describes the level of activation or intensity.

Recent advances in deep learning have significantly improved the performance of video-based emotion recognition systems. Visual cues such as facial expressions, head movements, and eye gaze are widely used to infer affective states [3]. However, relying solely on visual data is often insufficient due to challenges such as occlusion, illumination variance, and loss of subtle temporal dynamics. Audio, as another key modality, conveys complementary emotional information through prosody, tone, and rhythm [4]. The fusion of audio and visual modalities has thus become a promising direction for building robust and accurate CER systems.

Despite these developments, two critical challenges persist: (1) effectively modeling long-term dependencies across time, which are crucial for understanding emotional dynamics, and (2) integrating multimodal features in a lightweight and scalable architecture, particularly for real-time or resource-constrained applications. Traditional sequence models such as LSTM [5] and Transformer [6] have demonstrated strong temporal modeling capabilities, but often at the cost of high computational complexity.

To address these challenges, we propose LightMamba, a novel multimodal audio-visual framework for continuous emotion recognition. Our model extends a previous video-based system by incorporating audio features extracted via both VGGish and openSMILE toolkits, and introduces a lightweight early fusion mechanism. The concatenated multimodal features are processed by a four-layer Temporal Convolutional Network (TCN) followed by a Mamba block—a recently proposed selective state-space model designed for efficient long-sequence modeling [7]. Finally, a fully connected layer outputs the predicted valence-arousal values. The main contributions of this paper are as follows:

- We propose LightMamba, a novel and lightweight multimodal framework for continuous emotion recognition, which integrates visual and audio features using an early fusion strategy.
- We design a dual-branch audio feature extractor that combines VGGish and openSMILE embeddings to capture complementary acoustic cues for emotion modeling.
- We employ a hybrid temporal modeling architecture that leverages a TCN for local dynamics and the Mamba state-space model for efficient long-sequence learning, achieving superior performance on the Aff-Wild2 benchmark.

II. Related Work

A. Continuous Emotion Recognition

CER aims to predict human emotional states along continuous dimensions, most commonly valence and arousal, rather than discrete emotion categories. The valence-arousal model, first proposed by Russell [2], provides a two-dimensional representation that enables more nuanced emotional interpretation

This work was supported in part by the National Natural Science Foundation of China under Grant 62177029.

in real-world settings. Recent CER research has leveraged deep learning to improve temporal affect modeling from video [8], audio [4], or both modalities. Datasets like Aff-Wild2 have become the standard benchmark for CER due to their large-scale, in-the-wild annotations [9].

B. Multimodal Emotion Recognition

Multimodal emotion recognition [10] seeks to combine complementary cues from different input sources—typically facial expressions, vocal prosody, and body language—to improve robustness and accuracy. Early fusion strategies concatenate features from different modalities before learning [11], while late fusion or attention-based fusion combines modality-specific predictions or dynamically weighs features [12]. Audio has been shown to contribute significantly to emotion understanding, especially when visual signals are unreliable [1]. In this work, we follow the early fusion paradigm and utilize both VGGish [13] and openSMILE [14] extractors to obtain diverse acoustic features.

C. Temporal Modeling for Emotion Dynamics

Modeling the temporal evolution of emotions is critical for CER tasks. Recurrent models such as LSTM and GRU have been widely applied due to their ability to handle sequential data [5], but they suffer from vanishing gradients and computational inefficiency over long sequences. More recently, convolution-based models like TCNs have gained popularity for their parallelism and temporal receptive fields [15]. Transformer-based architectures have also been explored but often require significant computational resources [6].

Mamba, a recently proposed selective state-space model, offers a compelling alternative for long-sequence modeling. It achieves linear-time complexity with strong performance across sequence tasks by combining state-space recurrence with efficient parallel operations [7]. We adopt a hybrid architecture combining TCN and Mamba to benefit from both local temporal patterns and global sequence context in a lightweight manner.

III. Method

A. Framework Overview

As shown in Fig. 1, the proposed LightMamba framework is a multimodal architecture that integrates both visual and audio information for continuous emotion recognition. Specifically, visual features are extracted from cropped video frames using a Masked Autoencoder (MAE), while audio features are obtained using two parallel extractors: VGGish and openSMILE. The resulting features from all three branches are concatenated and fed into a four-layer TCN to capture local temporal patterns. A Mamba block is then used to model long-range dependencies efficiently, followed by a fully connected layer that predicts valence-arousal values for each frame.

B. Visual Feature Extraction

We adopt a MAE with a ViT-Large backbone to extract high-level visual features from facial video frames. Each frame is cropped to the facial region, resized to fit the MAE input, and normalized accordingly. To reduce overfitting and improve generalization, we load pretrained weights and apply partial fine-tuning: the patch embedding layer and the first 16 Transformer blocks are frozen, while the remaining layers are trained on the emotion recognition task.

The CLS token from each frame serves as a compact global descriptor of facial expression. The visual feature extraction process is the same as in our previous work Charon [26].

Let $f_t^v \in \mathbb{R}^{1024}$ denote the visual feature at time step t. These features are temporally aligned with the audio features and passed to the subsequent fusion and temporal modeling modules.

C. Audio Feature Extraction

To complement the visual modality, we incorporate a dual-branch audio feature extraction module that captures both low-level acoustic and high-level semantic representations from speech signals. The input audio is extracted from the same video as the visual stream and synchronized at the frame level.

VGGish Feature Branch: We utilize a pretrained VGGish model to extract high-level audio embeddings from the raw waveform. The audio signal is first converted into log mel-spectrograms and then fed into the VGGish network, which outputs 128-dimensional embeddings for each segment. These features encode semantic and prosodic characteristics such as tone, pitch, and rhythm that are relevant to emotional expression.

openSMILE Feature Branch: In parallel, we employ the openSMILE toolkit to extract hand-crafted low-level descriptors (LLDs), including pitch, energy, MFCCs, jitter, and shimmer. We use the widely adopted eGeMAPS feature set, resulting in a 62-dimensional feature vector per frame. These features are known to be effective for paralinguistic tasks and complement the deep features from VGGish.

Fusion and Alignment: The extracted VGGish and openSMILE features are concatenated to form a 190-dimensional audio representation:

$$f_t^a = \mathrm{Concat}(f_t^{\mathrm{VGGish}}, f_t^{\mathrm{SMILE}}) \in \mathbb{R}^{190}$$

Each audio feature vector f_t^a is temporally aligned with the corresponding visual feature f_t^v at the same time step t, ensuring synchronous multimodal fusion in the next stage.

D. Temporal Modeling

To capture both local and long-range temporal dynamics in emotional expressions, we adopt a two-stage modeling approach that combines a TCN and a Mamba-based state-space encoder.

Given a sequence of multimodal embeddings $\mathbf{F} = \{f_t\}_{t=1}^T$, where $f_t \in \mathbb{R}^{1214}$ represents the concatenated visual and audio

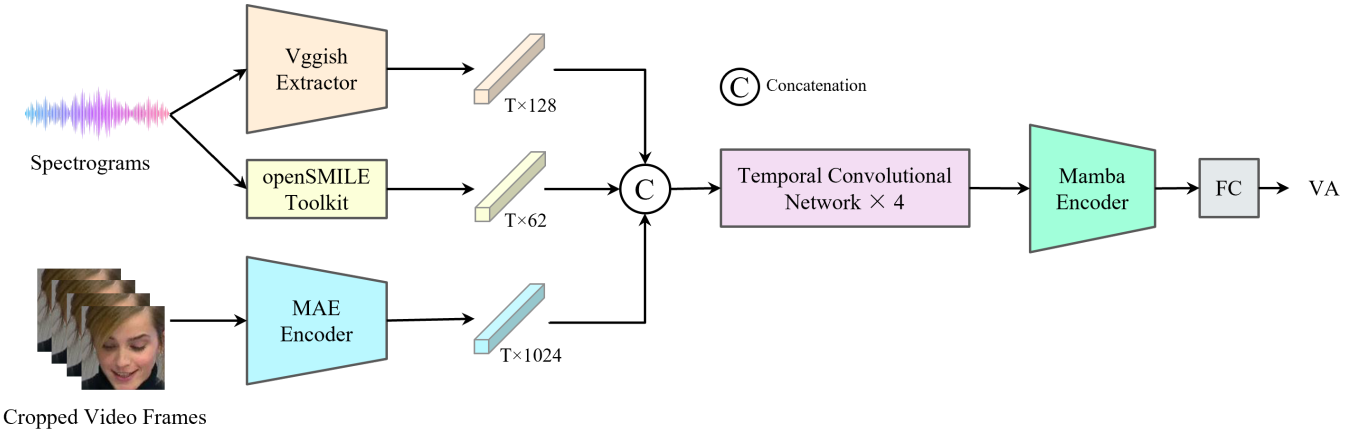

Fig. 1. Architecture of the proposed LightMamba model. Visual features are extracted using MAE, and audio features are obtained from VGGish and openSMILE. The fused features are processed by a four-layer TCN followed by four Mamba blocks for temporal modeling. The final regression head outputs valence and arousal values per frame.

features at time step t, we first apply a four-layer TCN to model short- and mid-range temporal dependencies:

$$\mathbf{G} = \mathrm{TCN}(\mathbf{F}), \quad \mathbf{G} = \{g_t\}_{t=1}^{T}, \quad g_t \in \mathbb{R}^D$$

Each TCN layer consists of dilated 1D convolutions with residual connections and ReLU activations. The dilation rate increases exponentially (1, 2, 4, 8), allowing the network to capture wider receptive fields without increasing model depth.

E. Mamba Encoder

The TCN-encoded sequence $\mathbf{G}$ is then passed to a four-layer Mamba encoder to further model long-term temporal relationships efficiently. Mamba is a recent state-space model that combines structured convolutional operations with dynamic state transitions, offering linear-time computation and strong sequence modeling capacity [7].

As shown in Fig. 2, each Mamba block processes the input sequence $X \in \mathbb{R}^{B\times L\times D}$ through two parallel branches:

- The first branch applies a linear projection, depthwise convolution, and SiLU activation, followed by a selective state-space model (SSM).
- The second branch applies a separate linear projection and SiLU activation, which acts as a dynamic gating signal.

The outputs of the two branches are fused via element-wise multiplication and projected back to the feature space. The output at each time step is given by:

$$h_t = W_2 \cdot (\mathrm{SSM}(\sigma(\mathrm{Conv}(W_1 X))) \odot \sigma(W_3 X))$$

where W_1, W_3, and W_2 are learnable projections, $\mathrm{Conv}(\cdot)$ is a depthwise convolution, σ is the SiLU activation, and $\odot$ denotes element-wise multiplication.

We stack four Mamba blocks to enhance modeling depth. The final output $\mathbf{H} = \{h_t\}_{t=1}^{T}$ is fed into a regression head to predict the frame-level valence and arousal values:

$$\hat{y}_t = \mathrm{FC}(h_t), \quad \hat{y}_t \in \mathbb{R}^2$$

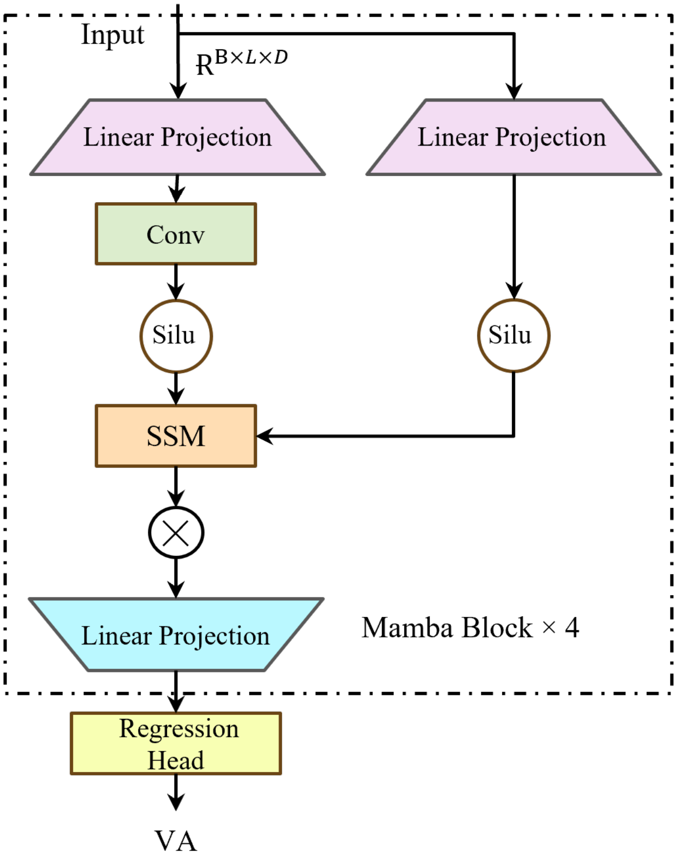

Fig. 2. Internal structure of the Mamba-based temporal module. The input is processed through two parallel branches: one with convolution and SSM, the other with gating. Their outputs are fused via element-wise multiplication and projected before regression.

This two-stage modeling design enables LightMamba to effectively capture both transient changes and global emotional trajectories in an efficient and scalable manner.

IV. Experiments and Results

A. Aff-Wild2 Dataset

We conduct all experiments on the Aff-Wild2 dataset [17], the largest and most diverse benchmark in affective computing. The dataset consists of 594 videos (approximately 3 million frames) featuring 584 unique subjects, including 16 videos with two annotated individuals. The videos are collected from YouTube and recorded under uncontrolled, in-the-wild conditions, making Aff-Wild2 highly representative of real-world affective behavior.

Each frame is continuously annotated for valence and arousal within the range of $[-1, 1]$, and the final labels are obtained by averaging the annotations from four expert raters. To ensure subject independence and avoid data leakage, the dataset is partitioned into training, validation, and test sets with no overlapping individuals.

Figure 3 illustrates the histogram distributions of valence and arousal across the dataset. Both distributions exhibit a positive skew, indicating that participants tend to express more positive and high-arousal emotions in the collected data.

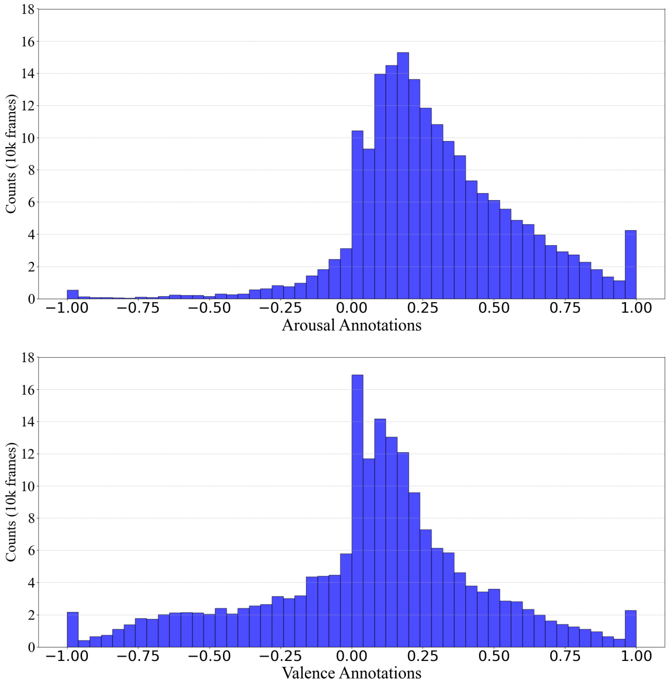

Fig. 3. Histogram of valence and arousal annotations in the Aff-Wild2 dataset.

For benchmarking, we use the baseline model provided by the 8th ABAW Challenge [22], which employs a ResNet-50 [18] backbone pretrained on ImageNet. A linear regression head is used to predict valence and arousal values, achieving CCC scores of 0.24 and 0.20, respectively.

Prior to feature extraction, all videos are segmented into individual frames and preprocessed using RetinaFace [16] to extract facial bounding boxes and five facial landmarks. The faces are then aligned using a similarity transformation based on eye, nose, and mouth landmarks, and resized to $224\times224\times3$. Pixel values are normalized to the range $[-1, 1]$. Frames with invalid annotations (e.g., values equal to -5) are discarded.

To handle missing or inconsistent labels, we adopt a refinement strategy inspired by Wang et al. [23], which leverages the correlation between discrete and continuous annotations to enforce temporal consistency and improve label quality.

B. Implementation Details

The proposed model is implemented using PyTorch and trained on two NVIDIA GeForce RTX 4090 GPUs (each with 24GB memory), ensuring sufficient computational resources for large-scale sequence modeling.

Considering the in-the-wild nature of the Aff-Wild2 dataset, we apply several data augmentation techniques to enhance model generalization. These include random rotation within ±10° to simulate natural head movements, horizontal flipping to increase invariance to facial orientation, and random cropping and scaling to reduce background distractions and emphasize facial regions. All input images are normalized using a mean of 0.389 and a standard deviation of 0.198, which stabilizes training and accelerates convergence.

The audio-visual features are aligned frame-wise and segmented into fixed-length windows of 300 frames with a stride of 200. This allows the model to learn from overlapping temporal contexts while maintaining computational efficiency. Each windowed sequence is processed through a four-layer TCN with exponentially increasing dilation rates to capture short- and mid-range temporal dependencies.

The output of the TCN is passed through a four-layer Mamba encoder, which models long-range temporal dependencies in a memory-efficient manner. Finally, a fully connected regression head predicts frame-level valence and arousal values.

Training is conducted for 50 epochs using the AdamW optimizer with an initial learning rate of 3×10^{-4} and a weight decay of 10^{-3}. A linear warm-up is applied over the first 5 epochs, and dropout with a rate of 0.3 is used throughout the temporal modules to prevent overfitting.

C. Evaluation Metrics

We evaluate the model performance using the Concordance Correlation Coefficient (CCC), a standard metric for continuous emotion prediction tasks. CCC measures the agreement between predicted and ground-truth sequences, considering both correlation and mean squared differences.

Given prediction x and ground truth y, the CCC is defined as:

$$\text{CCC} = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

where ρ is the Pearson correlation coefficient, μ_x, μ_y are the means, and σ_x, σ_y are the standard deviations of x and y, respectively.

We compute CCC separately for valence (CCC_v) and arousal (CCC_a), and report the average:

$$\text{CCC}_{avg} = \frac{\text{CCC}_v + \text{CCC}_a}{2}$$

Higher CCC values indicate better alignment between predictions and annotations, with a maximum score of 1 indicating perfect agreement.

D. Results and Comparison

1) Five-Fold Cross-Validation: To evaluate the robustness and generalization capability of LightMamba, we perform five-fold cross-validation on the Aff-Wild2 training set. Table I presents the CCC scores for valence, arousal, and their average across different folds.

TABLE I
COMPARISON OF CCC SCORES FOR VALENCE, AROUSAL, AND THEIR AVERAGE ACROSS TEST FOLDS (FOLD 1–5)

Fold	Valence (CCC)	Arousal (CCC)	Average
Fold 1	0.543	0.575	0.559
Fold 2	0.516	0.579	0.548
Fold 3	**0.610**	**0.664**	**0.637**
Fold 4	0.569	0.663	0.616
Fold 5	0.567	0.648	0.608
Baseline	0.240	0.200	0.220

The results demonstrate that LightMamba achieves consistently high performance across different folds, with an average CCC score of 0.637 on Fold 3—the highest among all. Compared to the baseline model, LightMamba improves the average CCC by over 40%, showing excellent generalization and robustness across varying data distributions.

2) Comparison with State-of-the-Art Methods: Table II compares our proposed LightMamba framework with top-performing submissions from the 8th ABAW [22] Challenge on the Aff-Wild2 dataset. LightMamba achieves the highest CCC scores across all metrics, demonstrating strong generalization and robustness in real-world affective computing scenarios.

TABLE II
COMPARISON WITH TOP SUBMISSIONS FROM THE 8TH ABAW CHALLENGE (2025) ON THE AFF-WILD2 DATASET.

Method	Valence (CCC)	Arousal (CCC)	Average
Baseline	0.240	0.200	0.220
CAS-MAIS	0.327	0.304	0.316
Charon [26]	0.504	0.412	0.458
AIWELL-UOC [25]	0.468	0.492	0.480
HSEmotion [24]	0.494	0.551	0.522
CtyunAI [21]	0.546	0.611	0.578
DeepAVER-CRIM [20]	0.561	0.620	0.590
USTC-IAT-United [19]	0.577	0.623	0.600
LightMamba (Ours)	**0.610**	**0.664**	**0.637**

Several leading methods adopted multimodal fusion strategies. For example, USTC-IAT-United [19] employed pre-trained ResNet and VGG encoders in conjunction with TCNs and cross-modal attention mechanisms. DeepAVER-CRIM [20] introduced a recursive joint cross-attention framework with gating modules to manage dynamic modality complementarity. These designs, although effective, often involve complex architectures or significant computational overhead.

In contrast, our proposed framework, LightMamba, focuses on lightweight and efficient multimodal integration. It combines high-level visual features extracted from a MAE with complementary audio features derived from VGGish and openSMILE. These modalities are fused early and passed through a hybrid temporal modeling structure that leverages both TCNs and the Mamba state-space model for efficient long-term sequence learning.

LightMamba is an extension of our previous work, Charon [26], which relied solely on visual inputs. By introducing multimodal fusion, LightMamba captures richer affective cues and achieves significant performance gains with only a modest increase in complexity. Unlike methods that depend on external pretraining datasets or handcrafted alignment modules, LightMamba directly fine-tunes the MAE on Aff-Wild2 and processes synchronized audio-visual sequences end-to-end.

Other teams, such as HSEmotion [24] and AIWELL-UOC [25], also explored multimodal fusion using handcrafted modules or multiple expert-designed branches. While these methods achieve competitive results, they often require extensive tuning. LightMamba, on the other hand, achieves state-of-the-art performance with a streamlined architecture, reduced memory usage, and better deployability, making it well-suited for practical affective computing scenarios.

3) Ablation Study: To evaluate the computational efficiency and modeling effectiveness of the Mamba block, we perform an ablation study by replacing it with a standard Transformer encoder of comparable hidden size and layer depth. The two models are compared in terms of total parameters, GPU memory usage, training time per epoch, and average CCC score on the validation set, as shown in Table III.

TABLE III
COMPARISON BETWEEN MAMBA AND TRANSFORMER-BASED TEMPORAL MODULES.

Method	Params	GPU Mem	Time/Epoch	CCC Avg
Transformer	85.7M	7.8 GB	23s	0.593
LightMamba	116.5M	4.8 GB	24s	0.637

Although LightMamba has a slightly higher parameter count due to its multimodal fusion components, its temporal modeling module—based on Mamba—achieves significantly lower GPU memory consumption (4.8 GB vs. 7.8 GB). This demonstrates that Mamba offers a more memory-efficient design, which is beneficial for deployment on resource-constrained platforms.

Furthermore, the per-epoch training time of both models is nearly identical (24s vs. 23s), indicating that Mamba's linear-time state-space modeling introduces no additional computational overhead compared to the Transformer. More importantly, LightMamba achieves a substantial performance gain of +4.4% in average CCC, validating the superior temporal modeling ability of Mamba over Transformer in continuous emotion recognition tasks.

Overall, these results highlight that Mamba not only improves predictive performance but also enhances resource efficiency, making LightMamba a practical and scalable framework for real-world affective computing applications.

V. Conclusion

In this paper, we proposed LightMamba, a lightweight and effective multimodal framework for continuous emotion recognition in the valence-arousal space. By integrating high-level visual features extracted via a pretrained MAE with complementary audio features obtained from both VGGish and openSMILE, the proposed model captures rich emotional representations from both facial expressions and vocal cues.

To efficiently model the temporal evolution of emotions, we employed a hybrid architecture consisting of a TCN for local dynamic encoding and a Mamba-based sequence model for long-range dependency learning. Extensive experiments on the Aff-Wild2 benchmark demonstrate that LightMamba achieves state-of-the-art performance while maintaining a significantly lower memory footprint and comparable training speed compared to Transformer-based counterparts.

The results confirm that Mamba is not only effective for emotion modeling but also offers practical advantages in terms of computational efficiency and deployment scalability. As such, LightMamba provides a promising solution for real-time affective computing applications, particularly in memory-constrained or latency-sensitive environments.

We believe that this work contributes a novel perspective to the field of affective computing by demonstrating the benefits of combining selective state-space modeling with multimodal fusion. In future work, we plan to further explore modality-aware fusion mechanisms and extend our model to handle more complex multimodal interaction scenarios, thereby advancing the development of emotionally intelligent human-computer interaction systems.

References

[1] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," Proc. 13th Annu. ACM Int. Conf. Multimedia, pp. 669–676, 2005.

[2] J. A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] R. E. Kaisler, M. M. Marin, and H. Leder, "Effects of Emotional Expressions, Gaze, and Head Orientation on Person Perception in Social Situations," *SAGE Open*, vol. 10, no. 3, 2020.

[4] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language—State-of-the-art and the challenge," Computer Speech & Language, vol. 27, no. 1, pp. 4–39, 2013.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, pp. 5998–6008, 2017.

[7] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2302.06665*, 2023.

[8] D. Kollias *et al.*, "Analysing affective behavior in the second ABAW2 competition," in *Proc. ICCV*, pp. 3652–3660, 2021.

[9] D. Kollias and S. Zafeiriou, "Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework," *arXiv preprint arXiv:2103.15792*, 2021.

[10] W. Zheng and B. Lu, "Multi-modal emotion recognition using deep neural networks," in Proc. Int. Conf. Affective Comput. Intell. Interact. (ACII), 2015.

[11] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Emotion recognition in the wild challenge 2014: Baseline, data and protocol," in Proc. ACM ICMI, pp. 461–466, 2014.

[12] A. Zadeh *et al.*, "Memory fusion network for multi-view sequential learning," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[13] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, pp. 131–135, 2017.

[14] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Computing*, 2010.

[15] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.

[16] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: single-shot multilevel face localisation in the wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5203–5212, 2020.

[17] D. Kollias and S. Zafeiriou, "Aff-Wild2: extending the Aff-Wild database for affect recognition," arXiv preprint arXiv:1811.07770, 2019. [Online].

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.

[19] J. Yu, Y. Wang, L. Wang, Y. Zheng, and S. Xu, "Interactive multimodal fusion with temporal modeling," arXiv preprint arXiv:2503.10523, 2025.

[20] R. G. Praveen and J. Alam, "Handling weak complementary relationships for audio-visual emotion recognition," unpublished, 2025.

[21] W. Zhou, C. Ling, and Z. Cai, "Emotion recognition with CLIP and sequential learning," arXiv preprint arXiv:2503.09929, 2025.

[22] D. Kollias, P. Tzirakis, A. S. Cowen, S. Zafeiriou, I. Kotsia, E. Granger, et al., "Advancements in affective and behavior analysis: The 8th ABAW workshop and competition," 2025.

[23] Z. Wang, J. Zheng, and F. Liu, "Improvement of continuous emotion recognition of temporal convolutional networks with incomplete labels," IET Image Process., vol. 18, pp. 914–925, 2024.

[24] A. V. Savchenko, "HSEmotion team at ABAW-8 competition: audio-visual ambivalence/hesitancy, emotional mimicry intensity and facial expression recognition," arXiv preprint arXiv:2503.10399, 2025.

[25] J. Cabacas-Maso, E. Ortega-Beltrán, I. Benito-Altamirano, and C. Ventura, "Enhancing facial expression recognition through dual-direction attention mixed feature networks and CLIP: application to 8th ABAW challenge," arXiv preprint arXiv:2503.12260, 2025.

[26] Y. Liang, Z. Wang, F. Liu, M. Liu, and Y. Yao, "Mamba-VA: a Mamba-based approach for continuous emotion recognition in valence-arousal space," arXiv preprint arXiv:2503.10104, 2025.