Enhancing Audio-Visual Affective Analysis via Bidirectional Dynamic Cross-Modal Attention and Masked Autoencoding

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*—Affective computing plays a crucial role in enhancing human-computer interaction and supporting mental health monitoring. However, two major challenges persist: (1) how to effectively extract discriminative features that capture subtle affective variations, and (2) how to model and align the distributional discrepancies of multimodal features across spatial and temporal domains to leverage complementary information from different modalities. In this paper, we propose a novel multimodal affective analysis framework to tackle these issues. To capture fine-grained affective cues in the visual modality, we introduce a hierarchical decoupling mechanism based on a Masked Autoencoder (MAE), eliminating the need for large-scale facial pre-training and aligning facial features with the affective semantic space. For effective multimodal fusion, we present a Bidirectional Dynamic Cross-Modal Attention (BDCA) mechanism that adaptively models interactions between modalities and enhances affective state estimation. The fused features are then encoded via a transformer, and Valence-Arousal(V-A) values are estimated through a fully connected layer. Experiments on the challenging Aff-Wild2 dataset demonstrate the effectiveness of our framework. Our method achieves Concordance Correlation Coefficients (CCC) of 0.605 and 0.658 for Valence and Arousal, respectively, outperforming existing state-of-the-art methods. These results highlight the potential of our approach in advancing audio-visual affective computing.**

***Index Terms*—Audio-Visual Affective Computing, Valence and Arousal, Cross-Modal Attention, Masked Autoencoder**

I. Introduction

Affective computing plays a vital role in understanding human mental states and supports applications such as mental health assessment (e.g., detection of anxiety and depression) and human-computer interaction [1]. Existing approaches to affective analysis can be categorized into discrete affective categorization and dimensional modeling. The former, based on Ekman's [2] theory of seven basic emotions , offers interpretability but fails to capture the continuous, evolving nature of real-world affective states. In contrast, dimensional models like the Valence-Arousal(V-A) framework [3] characterize valence as affective polarity and arousal as activation level, enabling fine-grained temporal modeling. This framework has become the foundation of dimensional affective analysis.

Despite progress, dimensional affective analysis still faces two key challenges: (1) extracting discriminative features that reflect subtle affective variations, and (2) aligning multimodal features with distinct spatial-temporal characteristics to fully leverage their complementarity.

To address these issues, research has shifted from unimodal to multimodal affective analysis. Early unimodal methods [4] suffer from noise sensitivity and limited robustness, as each modality alone lacks context. In contrast, multimodal learning [5] combines information from visual and acoustic sources, significantly improving affective state estimation.In this work, we propose a novel audio-visual affective analysis framework that enhances both feature extraction and cross-modal fusion. Traditional hand-crafted features [6] and CNN-based models [7] struggle with illumination changes and data dependency. Masked Autoencoders (MAE) [8] offer a promising self-supervised alternative, but most affective analysis applications [9] rely on large-scale facial expression pretraining. We introduce a hierarchical decoupled fine-tuning strategy using MAE that enables efficient, affective-aware visual representation learning without requiring massive labeled datasets.

The audio contains rich affective cues embedded in spectral, rhythmic, and prosodic patterns. However, the interaction between audio and visual modalities remains underexplored. To capture complementary affective cues, we integrate multiple audio features—VGGish [11], eGeMAPS, and MFCC [12].

Multimodal fusion remains challenging. Conventional early and late fusion methods [14] lack the ability to model deep inter-modal dependencies. Transformer-based attention mechanisms [13] improve long-range modeling but still rely on static fusion weights, unidirectional attention, and may lose temporal nuance.To overcome these limitations, we propose a Bidirectional Dynamic Cross-Modal Attention (BDCA) mechanism that dynamically models cross-modal interactions and reallocates attention weights based on affective context. BDCA enhances semantic alignment between modalities and improves robustness against temporal variations.

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

Our main contributions are summarized as follows:

- We propose a hierarchical decoupling strategy based on mask autoencoders for visual representation learning that avoids costly pre-training in facial expression.
- We propose a BDCA mechanism that adaptively captures intra- and inter-modal dependencies.
- Our method achieves state-of-the-art performance on the Aff-Wild2 dataset, with CCC scores of 0.605 for Valence and 0.658 for Arousal.

II. Related Work

Traditional discrete affective categorization methods, though highly interpretable, often struggle to capture nuanced and continuous affective states. In contrast, the V-A dimensional model [3] provides a continuous and fine-grained representation of affective dynamics and has become the mainstream paradigm in audio-visual affective analysis.

Early affective analysis methods in the visual domain were grounded in Ekman's theory of basic emotions [2] and focused on facial expression recognition. Advances in face detection led to improvements in visual-based affective modeling, but such methods are often sensitive to environmental noise, occlusions, lighting variations, and cultural factors. To mitigate these limitations, researchers began exploring multimodal approaches. Banse [15] emphasized the affective role of audio, while Cohen [16] introduced a multimodal framework using SVMs, demonstrating the complementary nature of visual and audio cues.

With the rise of deep learning, multimodal affective computing has made significant progress. Zheng [17] integrated visual and audio features under the V-A framework, while Poria [18] leveraged CNNs and RNNs to capture spatial and temporal dependencies. Despite promising results, many existing methods fail to model the intricate intra- and inter-modal relationships required for robust affective state estimation.

Facial expressions remain central to affective representation. Spatial regions (e.g., eyes, mouth) encode localized semantics, while temporal dynamics reflect evolving expressions. Early approaches [19] used 2D CNNs with LSTMs, while later works such as Temporal Convolutional Networks (TCNs) [20] captured multi-scale temporal structure. Fan [22] enhanced TCNs with spatial-temporal attention to better model dynamic facial cues. More recently, self-supervised methods like MAE [8] have shown promise in visual representation learning. However, applications in affective computing [9] often rely on extensive pretraining on facial datasets, increasing computational cost.

The audio modality offers affective cues through various low- and high-level descriptors. MFCCs capture timbre-related information, while eGeMAPS models prosodic and rhythmic properties [12]. Combining handcrafted and deep features has been shown to enhance robustness [21]. VGGish [11], trained on large-scale audio corpora, has become a standard for audio feature extraction. Deep learning approaches have explored both 1D CNNs for waveform modeling and 2D CNNs for spectrograms [14].

While multimodal affective analysis benefits from complementary audio-visual cues, challenges persist in effective fusion [10]. Feature heterogeneity and temporal misalignment between modalities remain major bottlenecks. Early fusion techniques—such as feature concatenation or element-wise operations—struggle with cross-modal synchronization [24]. To address this, Yu [23] proposed selective attention mechanisms that dynamically adjust modality importance, showing improved robustness.

However, most existing approaches still inadequately capture both intra-modal dynamics and cross-modal synergy, and lack mechanisms to adaptively reweight modalities based on affective context. These limitations motivate our proposed framework, which introduces a hierarchical MAE-based visual feature extractor and a BDCA module. BDCA allows deep semantic alignment across modalities and supports adaptive fusion guided by affective state variations. The proposed method is detailed in the following section.

III. Methodology

This section presents an overview of the proposed framework, followed by detailed descriptions of the MAE-based visual feature extractor and the BDCA fusion module.

A. Approach Overview

As shown in Figure 1, our framework integrates visual and audio features for dimensional affective analysis using a BDCA mechanism. The architecture consists of four main components. First, visual and audio features are extracted independently from video frames. Second, four-layer TCNs are used to model intra-modal temporal dependencies. Third, these temporally encoded features are fused via the BDCA module, which adaptively captures complementary and synergistic interactions across modalities. The fused representation is then processed by a Transformer encoder to capture deeper contextual dependencies, and finally, fully connected layers output Valence and Arousal predictions.

B. MAE-based Visual Feature Extraction

Feature extraction is crucial in affective analysis, as it directly influences the model's ability to represent fine-grained affective content. We propose a hierarchical fine-tuning strategy based on a MAE, illustrated in Figure 2. Our method leverages the MAE's general-purpose pretraining for visual reconstruction while adapting it to affective regression tasks.

Specifically, we freeze the lower layers responsible for low-level visual encoding to retain generalization across domains, and selectively fine-tune the upper layers that capture high-level semantics. Compared to traditional domain-specific pretraining paradigms, this strategy implicitly aligns fine-grained facial expression features with the emotion semantic space by decoupling layer-wise parameter updates, effectively avoiding the need for large-scale facial image pretraining.

During fine-tuning, a 224×224×3 image is divided into non-overlapping 16×16 patches. Each patch is projected into a 1024-dimensional embedding via a linear projection layer.

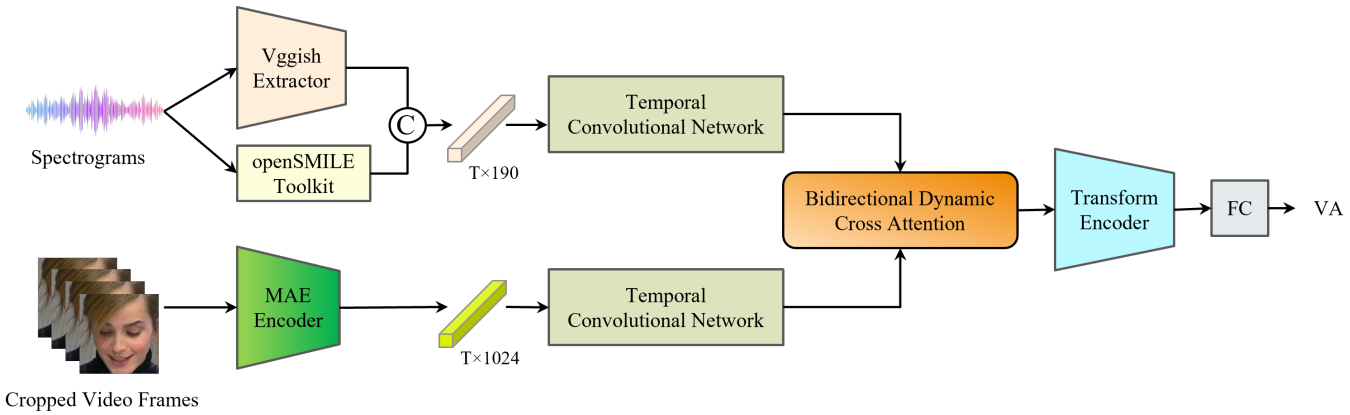

Fig. 1. Overall architecture of the proposed multimodal framework for dimensional affective analysis. Visual and Audio features are extracted, fused via BDCA, and refined by a Transformer encoder for valence-arousal estimation.

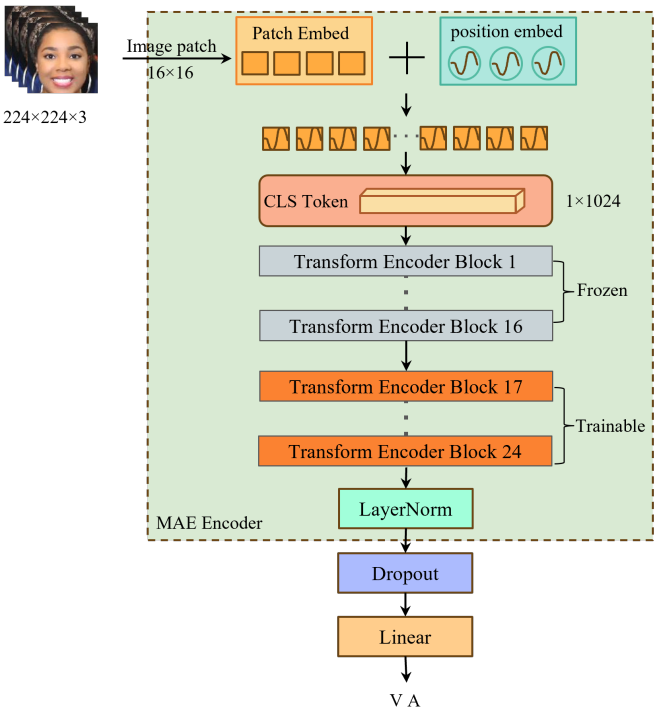

Fig. 2. Hierarchical fine-tuning strategy for the MAE encoder. Lower layers are frozen to retain general visual representations, while higher layers and the regression head are fine-tuned for V-A prediction.

To encode spatial relationships, fixed sinusoidal positional embeddings are added. A class token (CLS) aggregates global image information and is passed through the Transformer layers. The output of the CLS token is used as the global visual representation.

To balance efficiency and performance, a partial freezing strategy is adopted: the patch embedding layer and the first 16 Transformer blocks are frozen, while the last 8 blocks and the regression head are fine-tuned. This enables the model to preserve general visual features while focusing on affective-related semantics.

The resulting MAE encoder captures affective-relevant visual cues efficiently and transfers well to downstream tasks. Fine-tuning is performed on static, partially cropped video frames from the Aff-Wild2 training set (excluded from later model training). Only image data are used at this stage, audio and temporal signals are not introduced, to ensure that the extracted visual features remain modality-pure.

C. Audio Feature Extraction

We employ a multi-feature strategy to capture diverse affective cues from the audio modality. First, we use VGGish as the primary audio encoder to extract 128-dimensional representations. Trained on the large-scale VGGSound dataset, VGGish captures a broad spectrum of acoustic features and is used here as a fixed feature extractor without further fine-tuning.

In addition, we extract two widely used handcrafted audio descriptors using openSMILE [12]: MFCC and eGeMAPS. The MFCC features comprise 39 dimensions, including 13 base coefficients along with their first- and second-order derivatives, capturing the spectral and tonal variations that correlate with affective intensity. The eGeMAPS feature set consists of 23 dimensions, encoding affective-related prosodic and physiological parameters such as pitch variation, loudness, and speech rate.

D. Bidirectional Dynamic Cross-Modal Attention

To capture the synergistic and complementary relationships between audio and visual modalities, we propose a BDCA mechanism, as illustrated in Figure 3. BDCA employs two cross-attention modules—Visual-Driven Audio Attention (VDAA) and Audio-Driven Visual Attention (ADVA)—to enable bidirectional mapping between modalities. This design facilitates full modality interaction, enhances fine-grained feature representation, and adaptively reweights modality contributions to improve affective state estimation, overcoming the limitations of static fusion and unimodal modeling. Given a

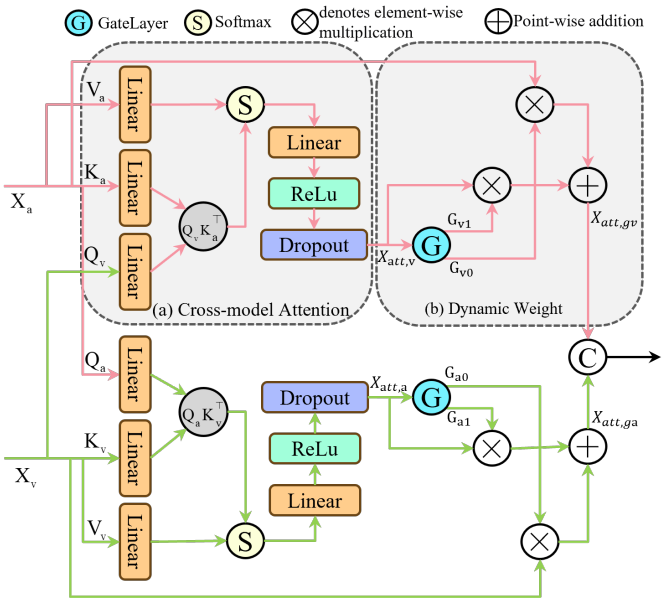

Fig. 3. Illustration of the BDCA fusion mechanism. Two cross-attention modules are constructed: VDAA and ADVA. A dynamic weighting module adaptively balances original and cross-attended features.

video sequence S with L frames, audio and visual streams are encoded by four-layer TCNs, producing:

$$X_v = \{x_v^1, x_v^2, \cdots, x_v^L\} \in \mathbb{R}^{d_v \times L},$$
$$X_a = \{x_a^1, x_a^2, \cdots, x_a^L\} \in \mathbb{R}^{d_a \times L}$$

where d_v and d_a are the dimensions of visual and audio features, respectively.

Each modality undergoes cross-attention. For VDAA, visual features X_v serve as queries, while audio features X_a provide keys and values:

$$A_v = \text{Softmax}\left(\frac{Q_v K_a^\top}{\sqrt{d_a}}\right) V_a \tag{1}$$

The attended features A_v are added to X_v, followed by ReLU activation:

$$X_{\text{att},v} = \text{ReLU}(A_v + X_v) \tag{2}$$

To adaptively regulate the importance of cross-attended and original features, we introduce a Dynamic Weight module, shown in Figure 3. A gating layer learns a weighted combination:

$$W_{\text{go},v} = X_{\text{att},v}^\top W_{\text{gl},v} \tag{3}$$

$$G_v = \frac{e^{W_{\text{go},v}/T}}{\sum_{K=1}^{K} e^{W_{\text{go},v}/T}} \tag{4}$$

where $W_{\text{gl},v} \in \mathbb{R}^{d_v \times 2}$ is a learnable parameter, and T is a temperature term (empirically set to 0.5) to smooth the softmax output [25]. K denotes the number of output units of the gating layer, which is 2. $G_v \in \mathbb{R}^{L \times 2}$ contains attention scores for the original and cross-attended features.

Each column of G_v is broadcasted to match feature dimensions, yielding G_{v0} and G_{v1}. These are used to reweight the original and attended features:

$$X_{\text{att},gv} = \text{ReLU}\left(X_v \otimes G_{v0} + X_{\text{att},v} \otimes G_{v1}\right) \tag{5}$$

where $\otimes$ denotes element-wise multiplication.

This dynamic fusion process allows the model to select or suppress cross-attended or original features based on their relevance. For highly complementary modalities, the attention weight for cross-attended features approaches 1, while that for the original approaches 0, and vice versa. This mechanism introduces soft regularization and improves generalization across varying affective contexts.

The ADVA module operates symmetrically, using audio features as queries and visual features as keys/values, producing $X_{\text{att},ga}$. The final joint multimodal representation is obtained by concatenating the gated features:

$$X_{\text{fused}} = [X_{\text{att},gv}; X_{\text{att},ga}] \tag{6}$$

This fused embedding serves as input to the Transformer encoder for final affect prediction.

IV. Experiments and Results

This section presents the experimental setup, evaluation results, and ablation studies, highlighting the effectiveness and robustness of the proposed framework in comparison with state-of-the-art methods.

A. Aff-wild2 Dataset

We conduct experiments on the Aff-Wild2 dataset [26], the largest and most diverse dataset in affective computing. It consists of 594 videos (approx. 3 million frames) from 584 subjects, including 16 videos with two annotated individuals. Videos are sourced from YouTube and captured in uncontrolled environments, making them well-suited for real-world affective analysis. Annotations for Valence and Arousal are continuously labeled within [-1, 1], averaged across four expert annotators. To ensure subject independence, the dataset is split into training, validation, and test subsets with no subject overlap.

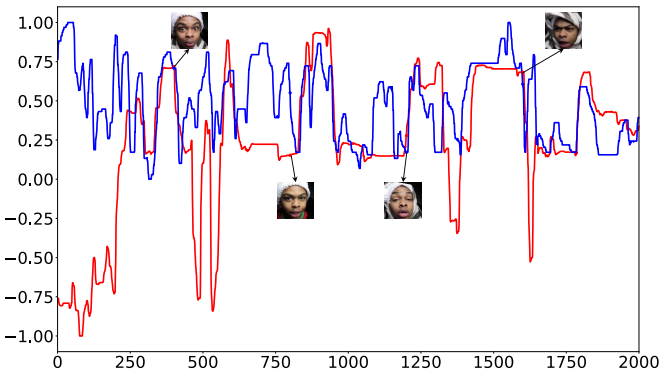

Fig. 4. Example trajectories of Valence and Arousal annotations over time for a sample video from Aff-Wild2.

Figure 4 illustrates an example of V-A annotation curves over time, highlighting challenges such as diverse expressions, rapid affect shifts, and facial occlusions. These characteristics underscore the dataset's complexity and realism.

Figure 5 shows the histogram distributions of Valence and Arousal. Both distributions are positively skewed, indicating a higher prevalence of positive valence and arousal levels among participants.

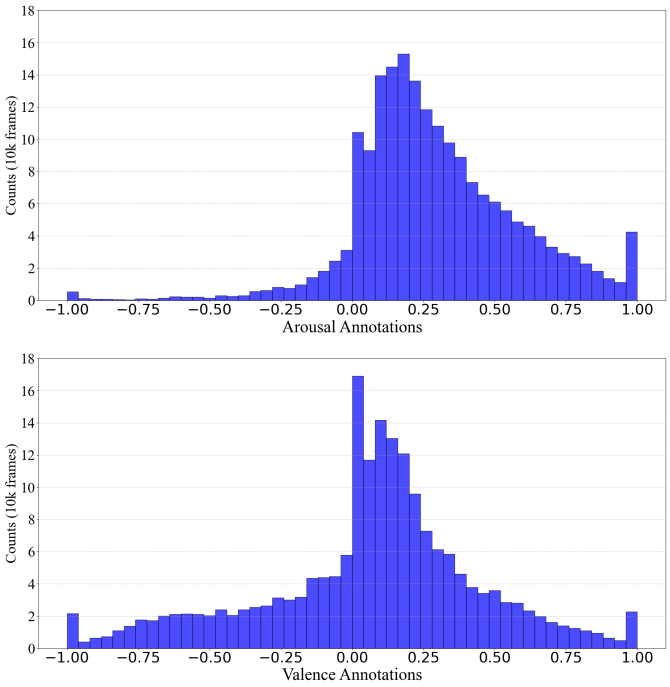

Fig. 5. Histogram of valence and arousal annotations in Aff-Wild2.

As a baseline, we benchmark against the 8th ABAW [36] Challenge model, which uses a ResNet-50 [27] backbone pretrained on ImageNet. A linear output layer predicts Valence and Arousal, achieving CCC scores of 0.24 and 0.20, respectively.

Before feature extraction, all videos are segmented into frames and processed with RetinaFace [28], which extracts facial bounding boxes and five landmarks. The faces are aligned via similarity transformation using eye, nose, and mouth landmarks, and resized to 224×224×3. Pixel intensities are normalized to [-1, 1]. Frames with invalid annotations (e.g., value = -5) are discarded. To address missing labels, we apply a refinement approach inspired by Wang [29], using the correlation between discrete and continuous labels to maintain temporal continuity.

B. Evaluation Metric

The dimensional sentiment recognition task involves continuous-valued sentiment prediction. The model design and optimization strategy must prioritize the stability of the regression task. The evaluation metric used is the average of the CCC for both Valence and Arousal:

$$L_{VA} = \frac{CCC_a + CCC_v}{2} \tag{7}$$

CCC quantifies the agreement between two time series (true and predicted values for all videos) by scaling their correlation coefficients with the mean square deviation. This approach penalizes predictions that are well-correlated with the true values but numerically skewed, in proportion to the deviation. CCC ranges from -1 to 1, where +1 indicates perfect agreement, and -1 indicates perfect disagreement. A higher CCC value indicates better alignment between the predicted and true values. CCC is defined as follows:

$$CCC = \frac{2S_x S_y \rho_{xy}}{S_x^2 + S_y^2 + (\bar{x} - \bar{y})^2} \tag{8}$$

Where ρ_{xy} is the Pearson Correlation Coefficient, S_x and S_y represent the variance of the predicted and true values of all the video Valence/Arousal dimensions, and $\bar{x}$ and $\bar{y}$ represent the mean of the predicted and true values, respectively.

C. Implementation Details

The proposed model is implemented using PyTorch and trained on two NVIDIA GeForce RTX 4090 GPUs (each with 24GB memory) to ensure computational efficiency and resource sufficiency.

Considering the in-the-wild nature of Aff-Wild2, we apply several data augmentation techniques to improve generalization. These include random rotation (±10°) to simulate natural head tilts, horizontal flipping to increase robustness, and random cropping and scaling to reduce background noise and emphasize facial regions. All input frames are normalized with a mean of 0.389 and a standard deviation of 0.198 to enhance training stability and accelerate convergence.

Both visual and audio modalities are segmented into overlapping temporal windows of 300 frames with a stride of 200, resulting in a 100-frame overlap between adjacent segments. This strategy captures temporal dependencies and affective transitions while increasing training samples for temporal modeling. Each window is independently passed through a four-layer TCNs to model intra-modal temporal dynamics, producing 128-dimensional representations per modality.

To prevent gradient instability and improve convergence, we adopt a learning rate warm-up strategy. The learning rate is linearly increased from 3×10^{-5} to 5×10^{-3} in the early training stages. Subsequently, we apply a cosine annealing learning rate scheduler (Cosine Annealing Warm Restarts) to dynamically adjust the learning rate across epochs, promoting smoother optimization and improved generalization.

D. Results and discussion

Table I presents the experimental results of our proposed framework on the V-A task test set. The CCC is used as the evaluation metric for both Valence and Arousal predictions. Fold 0 corresponds to the official test set, while folds 1–5 represent the results from five-fold cross-validation.

TABLE I
COMPARISON OF CCC SCORES FOR VALENCE, AROUSAL, AND THEIR AVERAGE ACROSS TEST FOLDS (FOLD 0–5)

Fold	Valence (CCC)	Arousal (CCC)	Average
Fold 0	0.582	0.646	0.614
Fold 1	0.562	0.613	0.588
Fold 2	0.556	0.605	0.581
Fold 3	0.605	0.658	**0.632**
Fold 4	0.539	0.689	0.614
Fold 5	0.582	0.650	0.616
Baseline	0.240	0.200	0.220

As shown in the table, our method significantly outperforms the baseline across all folds. These results demonstrate that the proposed MAE-based visual feature extractor effectively captures fine-grained affective cues, while the BDCA mechanism enables efficient and adaptive integration of audio-visual information. Together, these components substantially improve the model's performance on continuous affect estimation in real-world scenarios.

1) Comparison with State-of-the-Art Methods: Table II compares our BDCA framework with top submissions to the 8th ABAW Challenge on the Aff-Wild2 dataset. Our method achieves the highest CCC scores across all metrics, demonstrating strong generalization in real-world affective computing.

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 [35]	0.504	0.412	0.458
AIWELL-UOC [34]	0.468	0.492	0.480
HSEmotion [33]	0.494	0.551	0.522
CtyunAI [32]	0.546	0.611	0.578
DeepAVER-CRIM [31]	0.561	0.620	0.590
USTC-IAT-United [30]	0.577	0.623	0.600
BDCA(Ours)	**0.605**	**0.658**	**0.632**

Several leading methods employed multimodal fusion techniques. USTC-IAT-United [30] used pre-trained ResNet and VGG encoders with TCNs and cross-modal attention. DeepAVER-CRIM [31] extended a recursive joint cross-attention framework with gating mechanisms to handle varying modality complementarity. In contrast, our BDCA uses a simpler, non-recursive attention structure with dynamic weighting, offering better performance and lower complexity.

In terms of visual modeling, CtyunAI [32] fine-tuned CLIP on Aff-Wild2, while Charon [35] combined MAE, TCN, and Mamba for long-term temporal modeling. Unlike these approaches, we directly fine-tune MAE on the target task without external facial datasets, achieving superior results with fewer resources.

Other teams, such as HSEmotion [33] and AIWELL-UOC [34], incorporated additional modalities and handcrafted fusion designs. While effective, these methods may introduce greater architectural complexity. In general, BDCA achieves state-of-the-art performance with a streamlined design and a reduced reliance on large-scale pretraining.

2) Ablation Studies: To evaluate the individual contributions of key components in our framework, we conduct ablation studies on: (1) MAE fine-tuning, and (2) the BDCA fusion mechanism.

Effect of MAE Fine-Tuning. To evaluate the impact of visual representation learning strategies, we compare three configurations, as shown in Table III: (1) using MAE features without fine-tuning (MAE w/o FT), (2) using features from the Fuxi [9] model, which is pretrained on large-scale facial expression datasets and fine-tuned on Aff-Wild2, and (3) our proposed method, which fine-tunes the MAE encoder directly on Aff-Wild2 without relying on large-scale facial pretraining. In this experiment, only visual features were used, and the BDCA module was excluded to isolate the effect of visual encoding. These results demonstrate that self-supervised representation learning combined with task-specific fine-tuning is highly effective for in-the-wild affective analysis, even without access to large-scale affective-labeled facial datasets.

TABLE III
COMPARISON OF VISUAL FEATURE EXTRACTION METHODS ON AFF-WILD2.

Method	Valence (CCC)	Arousal (CCC)	Average
MAE w/o FT	0.348	0.419	0.384
Fuxi [9]	0.540	0.557	0.549
MAE	**0.554**	**0.595**	**0.575**

Effect of the BDCA Module. We further analyze the contribution of the BDCA module by comparing five configurations, as shown in Table IV. Using only the audio or visual modality leads to suboptimal performance, confirming the importance of multimodal integration. Replacing BDCA with simple concatenation of temporal audio-visual features yields slight improvements but still underperforms.

Introducing cross-modal attention (BDCA w/o DW) significantly boosts performance by enhancing inter-modal interactions. Adding the dynamic weighting mechanism further improves the average CCC by 1.94%, demonstrating its effectiveness in context-aware modality balancing. Overall, the full BDCA module achieves the best performance, highlighting its capacity to extract richer affective representations and adapt to complex affective dynamics.

V. CONCLUSION

This paper presents a novel and efficient multimodal framework for dimensional affective analysis, addressing key challenges such as dynamic temporal variation and the complexity

TABLE IV
IMPACT OF BDCA AND ITS COMPONENTS ON CCC PERFORMANCE.

Configuration	Valence (CCC)	Arousal (CCC)	Average
Baseline	0.240	0.200	0.220
Only Audio	0.212	0.334	0.273
Only Visual	0.554	0.595	0.575
Concat w/o BDCA	0.551	0.635	0.593
BDCA w/o DW	0.598	0.640	0.619
BDCA	**0.605**	**0.658**	**0.632**

of cross-modal fusion. To enhance feature quality and temporal modeling, we adopt a hierarchical MAE-based encoder with TCNs, enabling effective capture of fine-grained affective dynamics in both visual and audio modalities. Furthermore, the proposed BDCA module adaptively models inter-modal correlations and adjusts modality weights based on contextual reliability, improving robustness in real-world scenarios. A Transformer encoder is further employed to strengthen multi-scale temporal dependency modeling. Extensive experiments on the Aff-Wild2 dataset demonstrate that our approach outperforms state-of-the-art methods, confirming its ability to leverage audio-visual complementarity for accurate and robust affective state estimation. While the proposed model achieves strong performance, future work will explore deeper context-aware affect modeling and enhance cross-domain generalization to support broader deployment in human-computer interaction and mental health monitoring.

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] P. Ekman, "Basic emotions," in Handbook of Cognition and Emotion, T. Dalgleish and M. J. Power, Eds. New York, NY, USA: John Wiley and Sons, 1999, pp. 45–60.

[3] J. A. Russell, "A circumplex model of affect," J. Pers. Soc. Psychol., vol. 39, no. 6, pp. 1161–1178, 1980.

[4] S. Zhao, X. Yao, J. Yang, and others, "Affective image content analysis: two decades review and new perspectives," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6729–6751, 2021.

[5] D. Wang, T. Zhao, W. Yu, N. V. Chawla, and M. Jiang, "Deep multimodal complementarity learning," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 12, pp. 10213–10224, 2023.

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, pp. 886–893, 2005.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[8] K. He, X. Chen, S. Xie, et al., "Masked autoencoders are scalable vision learners," arXiv preprint arXiv:2111.06377, 2021.

[9] W. Zhang, B. Ma, F. Qiu, and Y. Ding, "Multi-modal facial affective analysis based on masked autoencoder," arXiv preprint arXiv:2303.10849, 2023.

[10] X. Li, G. Lu, J. Yan, and Z. Zhang, "A Survey of Dimensional Emotion Prediction by Multimodal Cues," Acta Automatica Sinica, vol. 44, no. 12, pp. 2142–2159, 2018. (in Chinese)

[11] S. Hershey, S. Chaudhuri, D. P. W. Ellis, et al., "CNN architectures for large-scale audio classification," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 131–135, 2017.

[12] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in Proc. 18th ACM Int. Conf. Multimedia, pp. 1459–1462, 2010.

[13] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 7676–7685, 2021.

[14] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," Pattern Recogn. Lett., vol. 146, pp. 1–7, 2021.

[15] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," Journal of Personality and Social Psychology, 1996.

[16] I. Cohen and I. Dagan, "Multimodal sentiment recognition in speech and text," in Proc. IEEE Int. Conf. Multimedia Expo (ICME), 2002.

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

[18] S. Poria and E. Cambria, "Multimodal emotion recognition in video: a deep learning approach," in Proc. Int. Conf. Affective Comput. Intell. Interact. (ACII), 2017.

[19] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Trans. Affect. Comput., vol. 2, pp. 92–105, 2011.

[20] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: a unified approach to action segmentation," in ECCV Workshops, pp. 47–54, 2016.

[21] L. Meng, Y. Liu, X. Liu, Z. Huang, W. Jiang, T. Zhang, et al., "Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 2344–2351, 2022.

[22] J. Fan, K. Zhang, Y. Huang, Y. Zhu, and B. Chen, "Parallel spatio-temporal attention-based TCN for multivariate time series prediction," Neural Comput. Appl., pp. 1–10, 2021.

[23] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, et al., "Multi-modal depression detection based on emotional audio and evaluation text," J. Affect. Disord., vol. 295, pp. 904–913, 2021.

[24] J. Q. Xiao and X. Luo, "A survey of sentiment analysis based on multi-modal information," in Proc. 2022 IEEE Asia-Pacific Conf. Image Process., Electron. Comput. (IPEC), pp. 712–715, 2022.

[25] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?," in Advances in Neural Information Processing Systems, vol. 32, 2019.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] R. G. Praveen and J. Alam, "Handling weak complementary relationships for audio-visual emotion recognition," arXiv preprint arXiv:2503.12261, 2025.

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

[33] 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.

[34] 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.

[35] 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.

[36] 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.