國立臺灣科技大學資訊工程系

112 學年度第 1 學期專題研究

總報告

基於面部情緒辨識技術的實況直播情緒共鳴應用

研究組員

B10915065 李昶勳

B10915036 余修辰

B10915059 許鎮承

B10915060 楊登傑

指導教授：戴碧如

中　華　民　國　　年　　月　　日

This is an English translation of the original project report, completed by Hsiu-Chien Yu (余修辰) under the supervision of Prof. Dai, Bi-Ru at NTUST. The signed Chinese cover page is attached for reference.

Abstract

This project aims to create resonance among viewers to enhance their viewing experience, while also allowing streamers to adjust their content based on audience reactions. To capture viewers' facial emotions in real-time and effectively, we improved upon a dynamic facial emotion recognition model by integrating a micro-expression recognition model to find the one with the best performance.

Preliminary results show that the model can detect viewers' facial emotions in real-time and effectively. We then use the Google Sheets API to consolidate the emotional information of all viewers and process this information to make the results easy to understand. Ultimately, this allows viewers to see the emotional reactions of others in real-time, enhancing the viewing experience.

Table of Contents

1. Introduction

As the pace of modern life accelerates, people often face significant stress, and some choose to watch live streams as a means of relief. However, compared to attending live performances in person, watching a live stream lacks the same on-site atmosphere.

If viewers could convey their true emotions over the internet, it could create resonance among them and greatly enhance the viewing effect. For streamers, they could also learn how to improve their content based on viewers' emotional feedback, thereby providing a better streaming experience.

To study whether this work can bring about a positive impact, we simulated a real-world scenario. We use facial emotion recognition technology to capture the genuine emotional reactions of the audience, consolidate all viewers' emotional information as quickly as possible, and present it in an easily understandable way to achieve a resonance effect.

To identify the most suitable facial emotion recognition model for our project, we trained micro-expression and dynamic facial emotion recognition models using various datasets. After comparing their respective strengths and weaknesses, we selected the model with the best performance.

2. Related Work

I. Dynamic Facial Emotion Recognition

In recent years, research on dynamic facial emotion recognition has grown (e.g., Former-DFER [1], IAL [2]). The key difference from static facial emotion recognition is that dynamic facial emotion recognition is video-based, while static facial emotion recognition is image-based. Additionally, videos better represent real emotions compared to images, which is why dynamic facial emotion technology has received considerable attention recently. Former-DFER was the first study to apply the Transformer to dynamic emotion recognition, and its results were significantly better than the commonly used CNNs. A common challenge in emotion recognition is that emotions can sometimes be subtle, leading to misclassification. IAL proposed a loss function to address this, but it did not lead to a significant overall performance improvement, largely due to insufficient data. Consequently, recent studies have adopted self-supervised learning, such as MAE-DFER [3], to tackle the data scarcity problem. MAE-DFER is currently the top-performing model on dynamic facial emotion recognition datasets (FERV39K [4], DFEW [5]), and our research is based on improving it.

II. Macro-Expression Emotion Recognition

For our application, we need to understand the audience's emotions quickly and in real-time, as their emotions are constantly changing while they watch. Furthermore, due to individual differences, people exhibit different emotional expressions to the same scene. Therefore, it is crucial to detect their emotions within a certain timeframe. Building on MAE-DFER, we also investigated the video-based MMEmotionRecognition [6] and the image-based POSTER-V2 [7] to compare the recognition capabilities of video and image models for various emotions.

III. Micro-Expression Emotion Recognition

The focus here differs from the above, which mainly concerns macro-expressions. Macro-expressions are defined as facial movements lasting longer than 0.5 seconds, whereas micro-expressions last for less than 0.5 seconds. Due to its wide range of applications, many have begun researching this area recently, but there has not been significant progress, also due to insufficient data. In our research, we adopted MMNET [8].

3. Methodology

I. Data Pre-processing

Before feeding the data into the model for prediction, we use mediapipe to accurately capture the viewers' faces.

Furthermore, for the input to the micro-expression recognition model, we employed a different approach from other studies. Previous research only used the Onset frame (the very first frame in a time period) and the Apex frame (the frame with the highest emotional intensity). We believe this method ignores too much emotional information, leading to model misclassification. Therefore, in this study, we chose to subtract the Onset frame from all frames within a period, take the average, and then feed it into the model for feature extraction.

II. Facial Emotion Recognition

- **Dynamic Emotion Recognition Model**

In our initial research, we used Former-DFER. Its architecture is shown below:

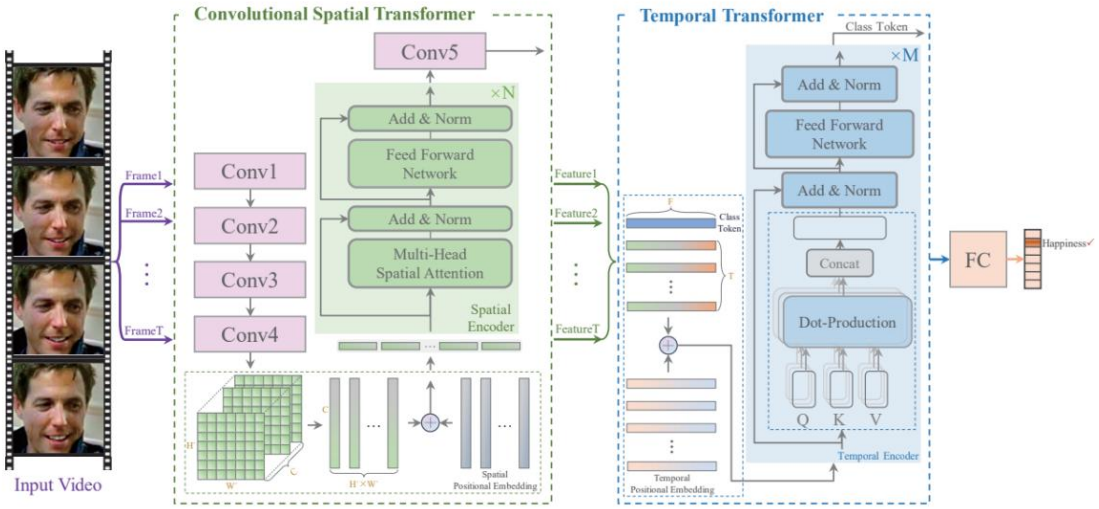

Fig 1. Former-DFER [1]

However, experiments showed particularly poor performance on "Disgust" and "Fear" because these two emotions are very rare in the dataset. Considering the dependency of supervised learning on data volume, we switched to a self-supervised model, MAE-DFER. This model is an improvement on VideoMAE [9], replacing ViT with LGI-Former to significantly reduce computational costs. Pre-trained on a large amount of unlabeled data, the model achieved an accuracy of 52.07% on FERV39K, making it the best-performing model to date.

The LGI-Former architecture and model comparison are shown below:

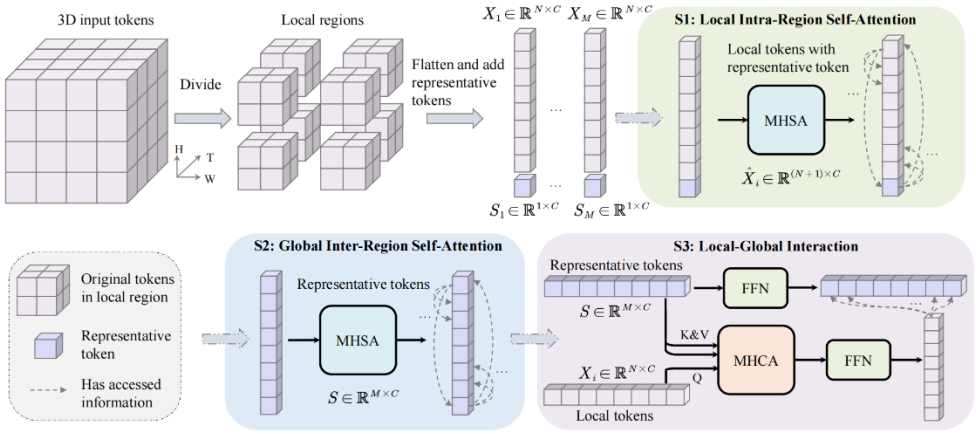

Fig 2. LGI-Former [3]

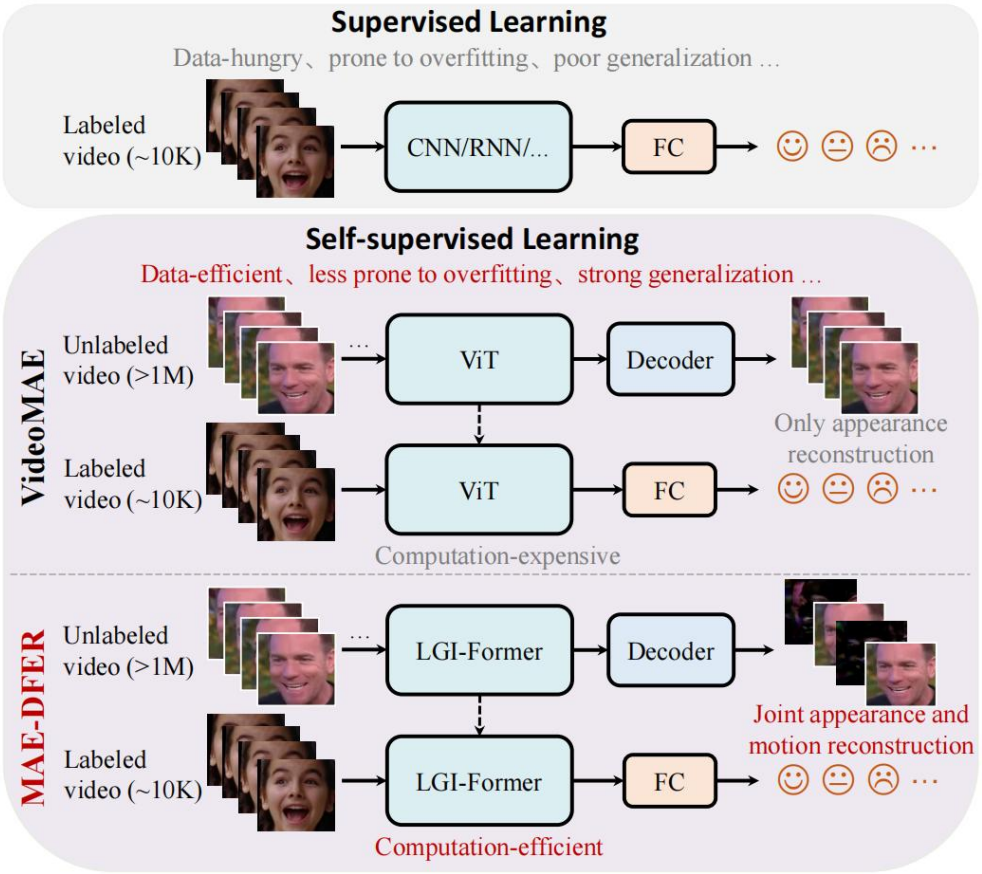

Fig 3. Comparison with VideoMAE [3]

During our experiments, we found that the model would misclassify results due to low emotional intensity, so we made corresponding improvements.

● Macro-Expression Recognition Model

Another video-based model is MMEmotionRecognition.

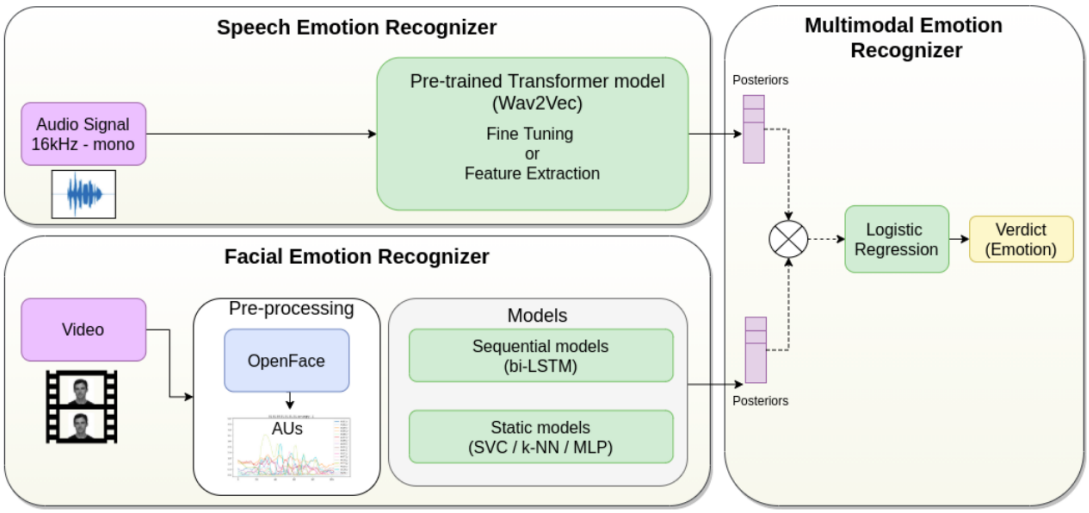

Fig 4. MMEmotionRecognition [6]

Initially, we extracted Action Units (AUs) from the video to record facial movements. We then used a static model, averaging the AUs from each frame for emotion analysis, but found that the accuracy and prediction speed did not meet our needs. This model also includes a separate speech emotion recognizer, which might be more suitable for streamers who frequently speak to the camera. Therefore, we believe it performs better on the streamer's end rather than for viewers who are simply watching.

● Static Emotion Recognition Model

The image-based model is POSTER-V2, its architecture is shown below:

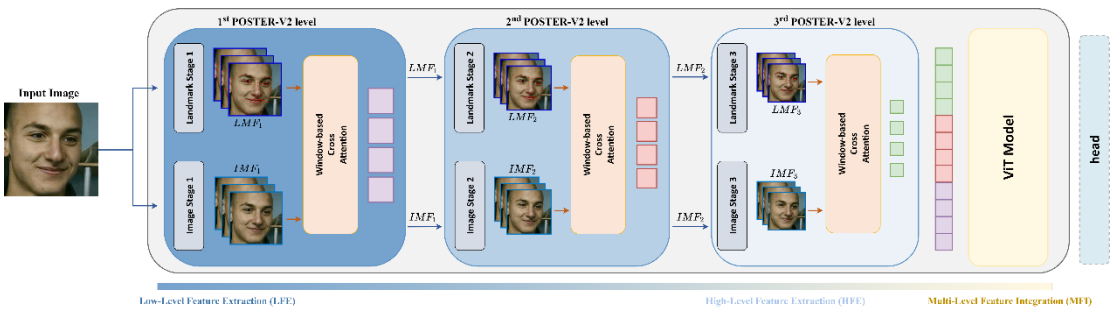

Fig 5. POSTER-V2 [7]

It processes images using facial feature detection and an image backbone, employing a dual-stream and cross-fusion method for multi-scale feature extraction. We found that combining multiple images to form a video can achieve a certain level of accuracy. Its accuracy on FERV39K is 49.79%, which, although slightly lower than MAE-DFER, performs slightly better on some specific emotions. Thus, we believe this model can be used to assist and verify the accuracy of other models.

- **Micro-Expression Recognition Model**

To address the issue of the model misclassifying results due to low emotional intensity, we observed that micro-expression models are more effective at detecting subtle facial changes. Therefore, we decided to combine a micro-expression model with a dynamic emotion recognition model. The model we chose is MMNET, its architecture is shown below:

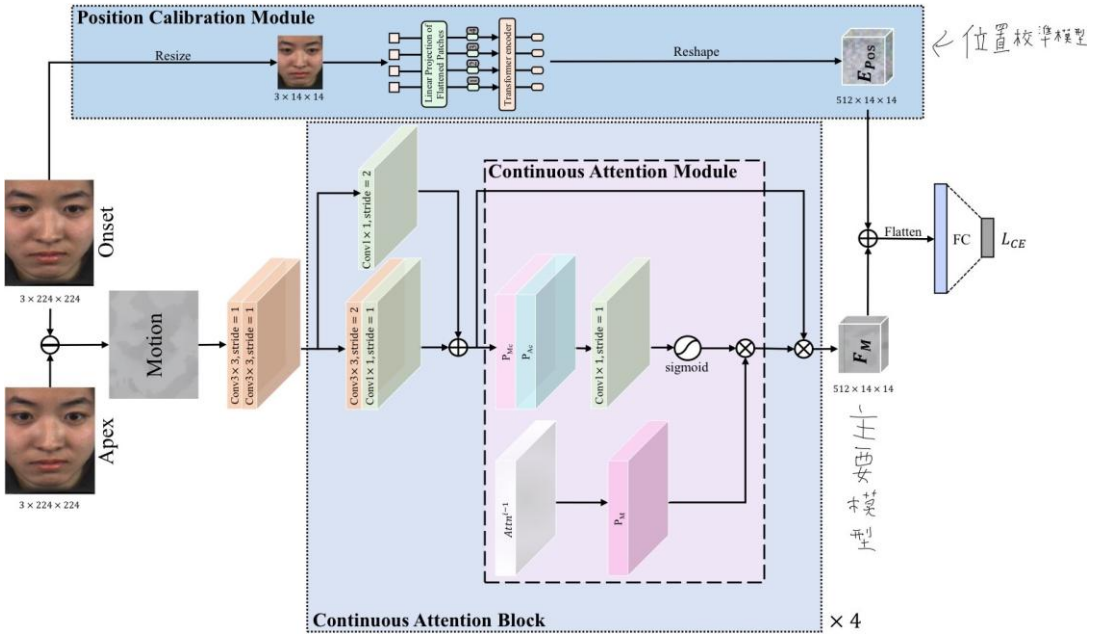

Fig 6. MMNET [8]

- Model Combination

To combine MMNET with MAE-DFER, we isolated the CA_Module from MMNET. We aimed to leverage the CA_Module's ability to extract muscle change features to enhance MAE-DFER's performance on subtle expressions. We added a fully connected layer after the original module to enable its combination with features extracted by LGI-Former. For feature combination, we tried two methods: concatenation and addition. The final result was obtained after feeding the combined features into a Fully Connected layer. The detailed architecture is as follows:

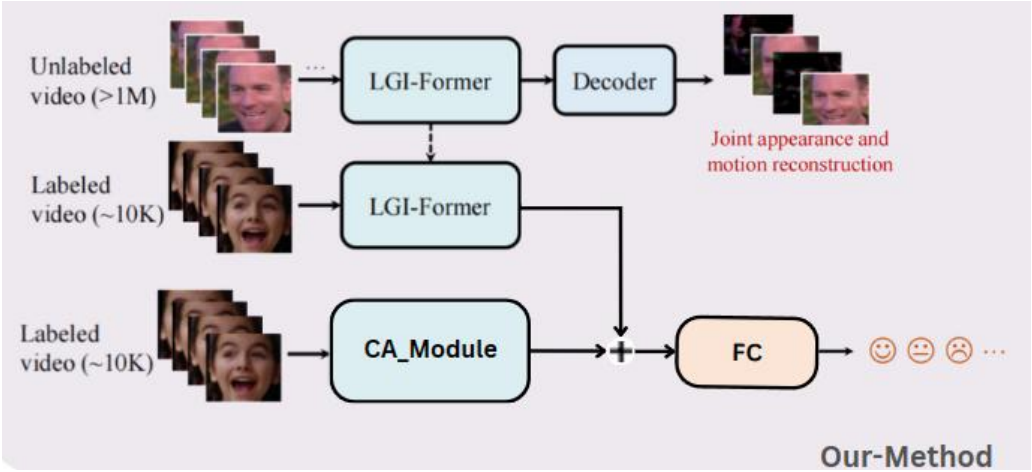

Fig 7. Our Method

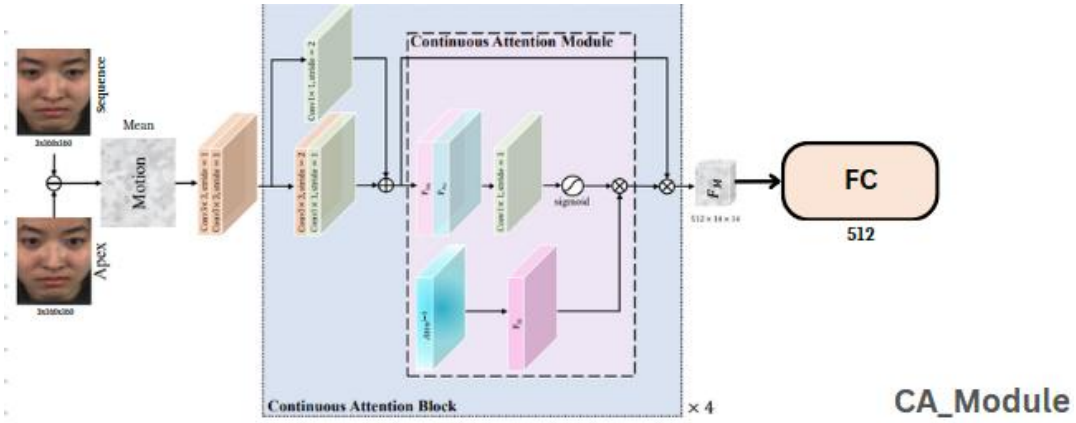

Fig 8. CA_Module

III. Consolidating Viewer Information

To achieve real-time reflection of each viewer's emotional changes, we utilized the Google Sheets API from the Google Cloud Platform with Python scripts. This allows us to send the emotions detected on the client-side to the cloud and then receive all viewers' emotions from the cloud, enabling transmission of emotional changes between clients and to the server.

IV. User Interface

We developed the user interface using tkinter due to its cross-platform compatibility and simplicity

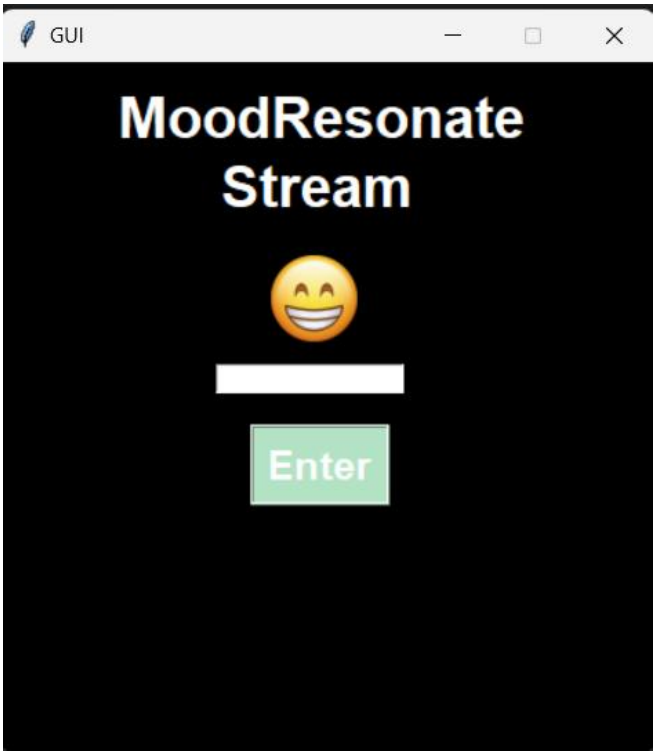

Fig 9. Entering a Username

Fig 10. Selecting a Live Channel

Fig 11. Emotion Recognition Function (Source: DFEW [5])

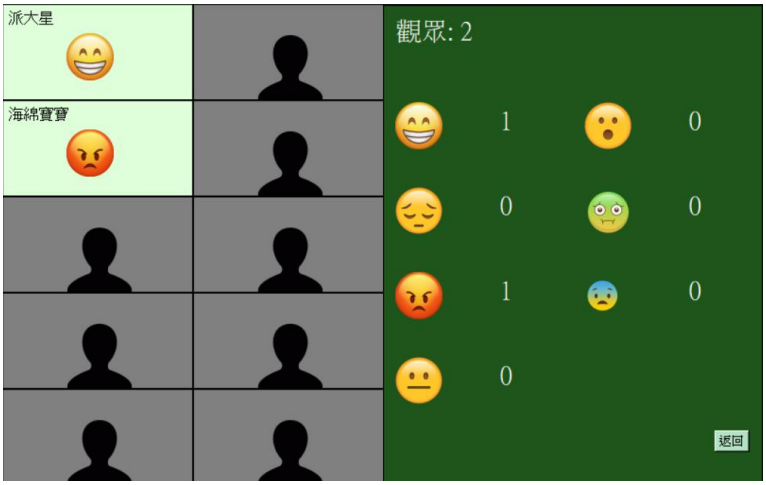

Fig 12. Synchronized Display of All Viewers' Emotions

In use, the left side shows each user's emotional reaction, while the right side displays a count of each emotion at that moment. This allows viewers to quickly understand the current emotions of the entire audience, fostering resonance and enhancing the viewing experience.

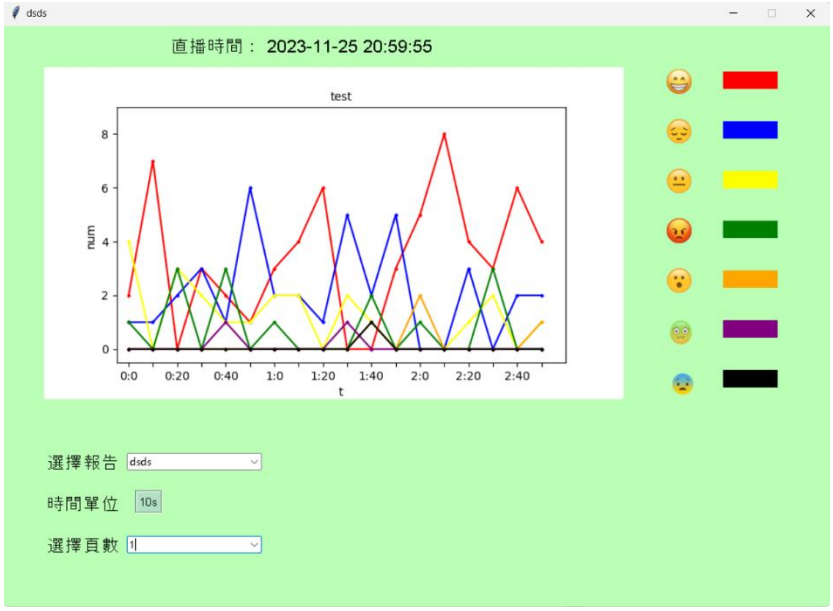

Fig 13. Displaying Channel History

After a live stream ends, we use a line chart to show the changes in viewer emotions over time. The purpose is to allow the streamer to understand the audience's emotional response to the content and improve it accordingly.

4. Results and Analysis

For our experiments, we used the training and testing datasets from FERV39K [4], which contains 38,935 videos across seven emotions: Happiness, Anger, Neutral, Sadness, Surprise, Disgust, and Fear.

First, to solve the problem of the model often misclassifying emotions due to low intensity, we tried the loss function proposed by IAL [2]. The formula is shown below:

$$\mathcal{L}_{IA} = -\log(P_{IA}), \tag{9}$$

$$P_{IA} = \frac{e^{x_t}}{e^{x_t} + e^{x_{max}}}. \tag{10}$$

Fig 14. Intensity-aware Loss[2]

x_t represents the model's output for the target class, and xmax represents the maximum output among the non-target classes. Our experimental results are as follows:

Loss function	**Accuracy**
Cross-Entropy	46.85%
Cross-Entropy + IAL	46.871%

Table 1. Performance of Former-DFER [1] on Ferv39k

The results show that this loss function did not effectively solve the problem. Therefore, we tried MMNET [8] alone. To compare it with dynamic facial emotion recognition methods, we also used FERV39K for training and testing. The results are as follows:

UAR(Unweighted Average Recall)	**WAR(Weighted Average Recall)**
25%	35%

Table 2. Results of the micro-expression model on Ferv39k

The results were very poor. We believe the poor performance is because the model is designed to focus on capturing subtle facial changes, which is different from macro-expression recognition models.

Dataset	Method	UAR	WAR
FERV39K	add	**42.30%**	**52.40%**
	concatenate	42.20%	52.30%

Table 3. Comparison of Feature Fusion Methods

The table below shows the results of combining the features extracted from the micro-expression model and the dynamic facial emotion recognition model using two different methods. The experimental results show that direct addition performs slightly better than concatenation.

We also tested the impact of each module of the micro-expression model on the overall model performance using the 'add' method.

CA_Module	PC_Module	#Params(M)	FLOPs(G)	UAR	WAR
×	×	85	50	42.09	51.82
✓	×	114	52	**42.30**	**52.40**
✓	✓	119	53	42.02	52.08

Table 4. Impact of Individual Modules on Model Performance

CA_Module stands for Continuous Attention Module, and PC_Module stands for Position Calibration Module.

The results show that adding the CA_Module slightly increases the computational load but improves the model's accuracy on low-intensity emotion data, with a WAR approaching SOTA performance. In the test where PC_Module was also added, the performance of using only CA_Module was better. We believe this is because LGI-Former already has a structure similar to PC_Module, and adding another PC_Module confused the facial position information, thereby reducing accuracy.

Method		Accuracy of Each Emotion(%)						
CA	**PC**	**Happy**	**Sad**	**Neutral**	**Anger**	**Surprise**	**Disgust**	**Fear**
×	×	72.51	53.41	61.18	50.44	26.18	**15.63**	**15.31**
✓	×	75.42	55.06	60.37	**50.84**	27.43	13.49	13.46
✓	✓	75.22	54.56	**61.44**	48.42	26.80	15.20	12.53
POSTER-V2		**78.07**	**56.71**	49.34	47.00	**28.84**	13.49	12.76

Table 5. Impact of Individual Modules on Accuracy for Each Emotion

The table above shows that POSTER_V2 [7] outperforms other models in predicting certain emotions, but its overall result is inferior to our MAE-DFER+MMNET (Ours) model.

Predict(emo) = Ours(emo) * w + POSTER-V2(emo) * (1 - w)

Fig 15. Model Combination Formula

To improve the accuracy for each emotion, we also tried combining MAE-DFER+MMNET (Ours) with POSTER-V2 [7] using weight ratios (w) of 0.8:0.2 and 0.6:0.4. The test on FERV39K yielded the following results:

Ours	**Poster**	**UAR**	**WAR**
✓	×	**42.30%**	**52.40%**
×	✓	40.89%	49.79%
0.8	0.2	40.82%	50.50%
0.6	0.4	40.83%	50.32%

Table 6. Impact of Different Weight Ratios on Model Performance

Using MAE-DFER+MMNET (Ours) alone still yields better performance than combining it with POSTER-V2. Even changing the weights did not result in a substantial improvement.

Method	**#Params(M)**	**FLOPs(G)**	**UAR**	**WAR**
Supervised methods				
C3D[11]	78	39	22.68	31.69
3D ResNet-18[12]	33	8	26.67	37.57
Former-DFER[1]	18	9	37.20	46.85
IAL[2]	19	10	35.82	48.54
POSTER-V2[7]	58	26	40.89	49.79
Self-supervised methods				
VideoMAE[9]	86	81	**43.33**	52.39
MAE-DFER[3]	85	50	42.09	51.82
MAE-DFER+MMNET(Ours)	114	52	42.30	**52.40**

Table 7. Results of Different Methods on FERV39K

Table 7 shows that self-supervised models significantly outperform supervised models. On our own hardware, compared to the original MAE-DFER [3], our model increased FLOPs by 2G but improved UAR and WAR by 0.21% and 0.58% respectively, bringing it closer to the SOTA model.

5. Conclusion and Future Work

Based on our current results, we can predict and display viewers' facial emotions on an interface in real-time with a certain degree of accuracy. This allows viewers to see others' reactions while watching a live stream, creating resonance and enhancing the viewing experience. Additionally, we have implemented a history function that uses a line chart to show viewers' emotional reactions over time, enabling streamers to understand audience feedback on their content and make improvements.

In the future, we hope to continue optimizing the model to improve its performance, especially on expressions with low emotional intensity. We also plan to enhance the user and streamer interfaces to improve the experience for both viewers and broadcasters

6. References

[1] Zhao, Zengqun, and Qingshan Liu. "Former-dfer: Dynamic facial expression recognition transformer." *Proceedings of the 29th ACM International Conference on Multimedia*. 2021.

[2] Li, Hanting, et al. "Intensity-aware loss for dynamic facial expression recognition in the wild." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 1. 2023.

[3] Sun, Licai, et al. "MAE-DFER: Efficient Masked Autoencoder for Self-supervised Dynamic Facial Expression Recognition." *arXiv preprint arXiv:2307.02227* (2023).

[4] Wang, Yan, et al. "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

[5] Jiang, Xingxun, et al. "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild." *Proceedings of the 28th ACM international conference on multimedia*. 2020.

[6] Luna-Jiménez, et al. "A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset." Applied Sciences, 12(1), 327.

[7] Mao, Jiawei, et al. "POSTER V2: A simpler and stronger facial expression recognition network." *arXiv preprint arXiv: 2301.12149* (2023).

[8] Li, Hanting, et al. "MMNet: Muscle motion-guided network for micro-expression recognition." *arXiv preprint arXiv:2201.05297* (2022).

[9] Tong, Zhan, et al. "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training." Advances in neural information processing systems 35 (2022): 10078-10093.

[10] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S., "CBAM: Convolutional Block Attention Module," arXiv preprint arXiv:1807.06521, 2018.

[11] Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

[12] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?." Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018.