Behavior-Based Tennis Match Outcome Prediction via Fusion of Class Balancing Strategy and Emotional Priors

Haozhe Bu[1,†], Yang Ma[1,2,†], Yunan Li[1,2,3,*] and Qiguang Miao[1,2,3]

[1] *School of Computer Science and Technology, Xidian University, China*
[2] *Xi'an Key Laboratory of Big Data and Intelligent Vision, China*
[3] *Key Laboratory of Smart Human-Computer Interaction and Wearable Technology of Shaanxi Province, China*

Abstract

This paper addresses the "Behavior-Based Emotion Recognitio" task in the IJCAI 2025 Micro-Gesture Hidden Emotion Understanding Challenge (MiGA) by proposing a tennis match outcome prediction method that integrates an emotional prior module, dual-channel image modeling, and a class balancing strategy. The task involves predicting the match outcome (win or loss) from post-match interview videos, where challenges include complex emotional expressions, subtle behavioral cues, and severe class imbalance.

To address the controlled emotional expressions and occluded facial cues, we first introduce a DeepFace-based emotion recognition module that performs frame-by-frame facial analysis to extract multidimensional emotional distributions and generate semantic tendency scores as prior information for the main model. On this basis, a dual-channel image feature extraction structure is designed, modeling global behaviors and local expressions separately using a shared ResNet34 backbone, thus enhancing the model's perception and discrimination of key visual cues. To mitigate the severe imbalance—over 70% of the training samples belong to the "win" class—we propose an intra-class partitioning strategy. The majority class is split into two subsets, each paired with the minority class to train two structurally identical but parameter-independent image classifiers, thereby implicitly achieving class balance at the data level. Finally, the two image classifiers are integrated with the emotional prior module, and the final prediction is determined via majority voting, significantly improving system robustness and generalization.

Experimental results demonstrate the proposed method's effectiveness in addressing challenges such as class bias, emotional concealment, and sparse signals, validating its potential in behavior modeling and complex emotion understanding scenarios.

Keywords

Behavior analysis, Emotion understanding, Class imbalance, Emotion prior, Video classification

1. Introduction

With the continuous advancement of artificial intelligence, video-based human behavior analysis and emotion understanding have become pivotal research areas in computer vision, showing great value in real-world applications such as sports analytics, psychological assessment, educational feedback, and media analysis [1] [2] [3]. Especially in unstructured scenarios where individuals' emotional expressions are controlled, facial cues are weak, and context is complex, accurately identifying hidden emotions to infer individuals' true states poses a critical challenge.

The IJCAI 2025 Micro-Gesture Hidden Emotion Understanding Challenge (MiGA) introduces a highly realistic task setting—predicting athletes' match outcomes (win or loss) based on post-match interview videos. Essentially a behavior-driven video classification problem, the task must be solved under significant uncertainty and data bias, making it extremely challenging. Specifically, the task presents two major difficulties: first, there is significant individual variation in emotional expression, and some

MiGA@IJCAI25: International IJCAI Workshop on 3rd Human Behavior Analysis for Emotion Understanding, August 29, 2025, Guangzhou, China.

[*] Corresponding author.
[†] These authors contributed equally to this work.

22009200599@stu.xidian.edu.cn (H. Bu); 22031212217@stu.xidian.edu.cn (Y. Ma); yunanli@xidian.edu.cn (Y. Li); qgmiao@xidian.edu.cn (Q. Miao)
https://orcid.org/0009-0005-6403-495X (H. Bu); https://orcid.org/0009-0006-8825-7311 (Y. Ma); https://orcid.org/0000-0001-7316-4354 (Y. Li); https://orcid.org/0000-0002-2872-388X (Q. Miao)

losing athletes still maintain a positive demeanor, making it difficult to directly correlate facial emotion signals with match outcomes. Second, the training set is severely imbalanced, with over 70% of the samples labeled as “win”, which easily leads to model overfitting and bias during training.

To tackle these challenges, this paper proposes an integrated prediction method that combines an emotional prior module, a dual-channel image feature extraction mechanism, and an intra-class partitioning strategy, aiming to improve the model’s capacity to identify latent emotions. The overall architecture is illustrated in Figure 1. Specifically, we extract both full-body and facial image frames from interview videos to construct two separate channels for learning behavioral and facial expression features, respectively. Temporal features from both channels are extracted using a shared ResNet34 backbone. A DeepFace-based emotion recognition module [4] [5] [6] [7] is then incorporated to perform frame-level facial analysis, generating multi-class emotional distributions and semantic prior scores as auxiliary judgment criteria. To address class imbalance, the majority “win” class samples are divided into multiple subsets, each combined with the complete minority “loss” class to train structurally identical but independently parameterized sub-models. Finally, these two image classifiers are integrated with the emotion prior module, and a majority voting mechanism is employed to produce the final prediction, thereby enhancing the system’s stability and generalizability.

The main contributions of this paper are as follows:

- We introduce a DeepFace-based emotional prior module to improve the recognition of complex and implicit emotional states;

- We construct a dual-channel image feature modeling framework incorporating both full-body behavior and facial expressions, and apply ensemble voting to achieve robust classification under high uncertainty and noise;

- We propose an intra-class partitioning and dual-model training strategy to effectively alleviate overfitting and prediction bias caused by severe class imbalance;

Experimental results demonstrate that our method performs well in modeling realistic complex interactions between behavior and emotion, validating its potential in multi-source information fusion and imbalanced data learning scenarios.

2. Related Work

In recent years, video-based behavior analysis and emotion recognition have become prominent research areas in computer vision, with wide applications in human-computer interaction, intelligent surveillance, and affective computing. In unstructured environments, emotional expressions are often subtle and complex, posing significant challenges to model accuracy and decision stability. Three major obstacles commonly arise: (1) class imbalance in training data, (2) implicit and ambiguous emotional signals, and (3) instability in model decisions under noisy conditions. To address these issues, prior studies have explored solutions from three main perspectives: class balancing techniques, emotional prior integration, and ensemble modeling strategies.

2.1. Class Imbalance Modeling Strategies

To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) [8] generates synthetic samples for minority classes, while Cui et al. [9] proposed a Class-Balanced Loss to reweight training data based on the “effective number” of samples. Kang et al. [10] further refined sample selection to reduce redundancy and focus on discriminative instances. However, these approaches often face trade-offs between maintaining data diversity and avoiding overfitting. Recent studies have explored implicit balancing strategies that preserve data integrity while improving minority class learning, motivating our use of intra-class partitioning in model design.

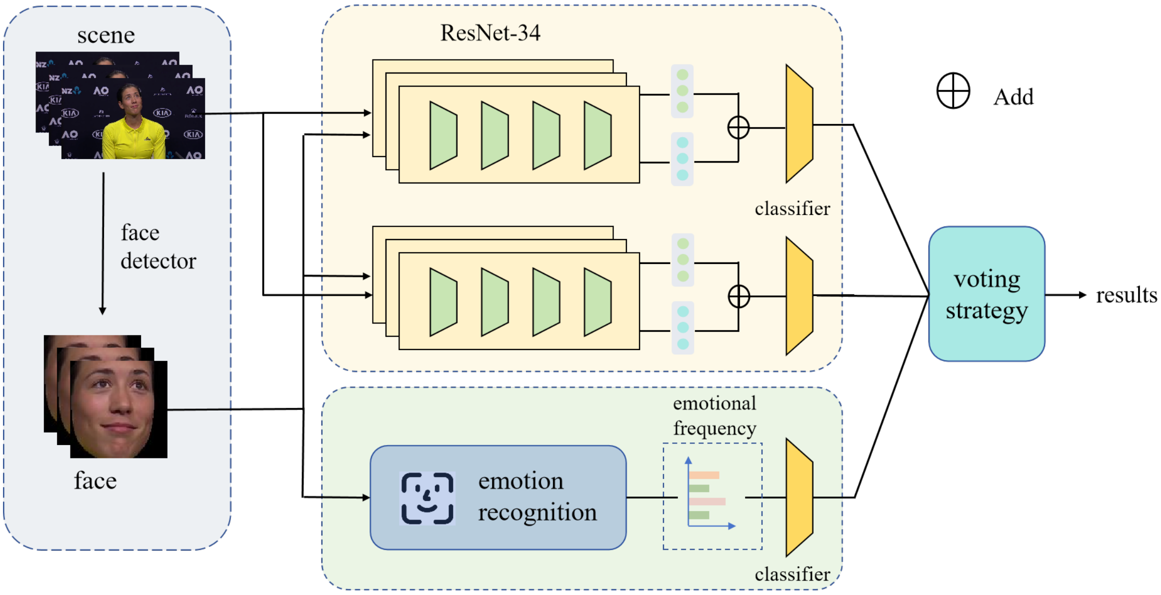

Figure 1: The architecture of the proposed method. At the top, two classifiers are trained using an intra-class partitioning and dual-model collaboration strategy. Each classifier adopts a dual-channel image feature extraction module to separately encode scene and facial features, which are then fused. At the bottom, the emotion-aware decision support module provides prior knowledge based on facial emotion recognition. Finally, a voting module integrates predictions from all components to determine the final classification result.

2.2. Incorporation of Emotional Priors

Emotion-related cues have been leveraged to support behavior recognition, particularly when primary features are ambiguous. Zhao et al. [11] addressed expression ambiguity by introducing RAF-DB and a deep learning architecture tailored for emotion recognition in the wild. Other works utilized facial Action Units (AUs) for fine-grained labeling [12], though these methods often depend on clean, high-resolution inputs. Building on this, some recent efforts have explored deep-learning-based emotion scoring as a soft auxiliary signal, offering greater robustness under occlusion and noise. Our work adopts a similar philosophy by incorporating facial emotion priors into the decision pipeline.

2.3. Structural Ensemble Mechanisms in Image-Based Behavior Modeling

Ensemble learning has proven effective in improving model robustness and generalization in expression recognition tasks. Hasani and Mahoor [13] proposed a two-stream network combining residual and attention mechanisms, while Corneanu et al. [14] highlighted the importance of structural diversity in ensemble systems. These findings inform our design of a structurally redundant yet independently trained ensemble framework, which integrates classification outputs and auxiliary emotion cues via a simple yet effective majority voting mechanism.

3. Methodology

The overall architecture is illustrated in Figure 1. Through the coordinated operation of multiple modules, the framework systematically addresses challenges posed by complex emotional expressions, weak behavioral cues, and imbalanced class distributions. The key components are described as follows.

3.1. Data Preprocessing

The preprocessing pipeline consists of three stages: frame sampling, temporal segmentation, and multi-scale feature construction. To reduce redundancy from the original 25 fps videos, we adopt sparse sampling at 2.5 fps, preserving temporal context while lowering computational cost. The downsampled video is split into 128-frame windows—overlapping or not—as training samples, each inheriting the original label. To enhance temporal representation, each segment is divided into 16 intervals with one frame randomly selected per interval. Additionally, we extract two crops from each sampled frame: a global full-body image and a localized face region detected via OpenFace. This dual-view strategy supports the later dual-channel modeling and boosts sensitivity to both macro behaviors and fine-grained emotional cues.

3.2. Emotional Prior Module

In post-match sports interview scenarios, emotional states often correlate with match outcomes: winners tend to exhibit joy, excitement, or relaxation, while losers may show sadness, tension, or suppression. This emotional-outcome association serves as a natural prior for classification.

Let f_i denote the frequency of the i-th emotion in the video, and w_i^+, w_i^- be its positive and negative weights. The total positive and negative tendency scores are calculated as:

$$\text{Positive Score} = \sum_i w_i^+ \cdot f_i \tag{1}$$

$$\text{Negative Score} = \sum_i w_i^- \cdot f_i \tag{2}$$

If the Positive Score exceeds the Negative Score, the emotional prior module predicts a win; otherwise, a loss.

This emotional prior provides semantically rich external knowledge, offering supplementary judgment signals that improve the system's discrimination capability and generalization performance.

3.3. Dual-Channel Image Modeling Module

While emotional priors often correlate with match outcomes, there are scenarios where this relationship is disrupted due to emotional masking or individual differences. For example, some athletes may maintain a positive appearance after losing due to professionalism, while others may appear emotionally subdued despite winning. To enhance the model's capability in recognizing such latent behavioral signals, we design a dual-channel image modeling module that jointly processes full-body and facial images using a shared ResNet34 backbone. Each channel extracts temporal features independently, with shared weights to reduce model complexity and enforce feature consistency.

The *scene channel* encodes macro-level visual cues such as posture, movement rhythm, and body language. The *face channel* focuses on micro-level expression features and subtle emotional variations.

Let $\mathbf{f}_{\text{scene}} \in \mathbb{R}^d$ and $\mathbf{f}_{\text{face}} \in \mathbb{R}^d$ denote the feature vectors extracted from the scene and face channels, respectively. These features are fused via element-wise addition:

$$\mathbf{f}_{\text{fused}} = \mathbf{f}_{\text{scene}} + \mathbf{f}_{\text{face}} \tag{3}$$

The fused representation is then passed through a linear classifier for binary prediction:

$$\hat{y} = W \cdot \mathbf{f}_{\text{fused}} + b \tag{4}$$

where $W \in \mathbb{R}^{1\times d}$ and $b \in \mathbb{R}$ are the parameters of the output layer, and $\hat{y} \in \mathbb{R}$ denotes the predicted score for the win/loss classification.

This design effectively integrates both global and local visual cues, enhancing the model's robustness and discriminative power under complex emotional masking conditions.

3.4. Intra-Class Partitioning and Dual-Model Training

In the iMiGUE dataset used for the MiGA challenge, a severe class imbalance exists—“win” samples far outnumber “loss” samples—leading to training bias and degraded generalization performance. To address this issue, we propose an intra-class partitioning strategy at the data level, combined with a dual-model training scheme. Specifically, the majority class (“win”) is randomly divided into two disjoint subsets of approximately equal size. Each subset is then paired with the complete “loss” class to form two balanced training datasets. Two independent ResNet34 models are trained on these datasets, which we refer to as BalanceNet A and BalanceNet B.

This strategy avoids the distributional distortion introduced by traditional oversampling or undersampling techniques. Moreover, by training on distinct data partitions, it introduces model diversity, which enhances the effectiveness of ensemble learning and reduces the risk of overfitting. In summary, the proposed strategy mitigates class imbalance at both the data and model levels, significantly improving fairness and generalization across the framework.

3.5. Multi-Model Ensemble via Voting Mechanism

To further enhance robustness, we adopt a majority-voting based multi-model ensemble consisting of three components: BalanceNet A, trained on “win” subset A and the full “loss” class; BalanceNet B, trained on “win” subset B and the full “loss” class; and EmotionScoreNet, an emotional prior module that generates predictions based on DeepFace emotion distribution scores.

Each model independently predicts the match outcome for a given video. The final decision is determined through majority voting: if at least two out of the three models agree on a class label, that label is taken as the final prediction.

This ensemble approach combines the complementary strengths of behavioral modeling and semantic emotion priors , significantly enhancing the system’s robustness in noisy or ambiguous conditions. Experimental results confirm that the ensemble consistently outperforms any individual model, demonstrating the effectiveness of the proposed architecture in complex emotion-behavior modeling scenarios.

4. Experiments and Results Analysis

4.1. Dataset and Evaluation Protocol

iMiGUE dataset [15]is utilized for the experiments, comprising 359 post-match interview videos collected from 72 professional tennis players. It adopts a cross-subject split strategy, where videos from 37 subjects totaling 245 samples are used for training, 10 videos are reserved for validation, and the remaining 104 videos from 35 distinct subjects constitute the test set.

The task is formulated as a binary classification problem. A label of 1 indicates a “win”, and 0 indicates a “loss”. Classification accuracy is adopted as the primary evaluation metric.

4.2. Experimental Setup

All experiments are conducted using the PyTorch framework. Input images are uniformly resized to 224×224 pixels. The classification head comprises two fully connected layers and is optimized with the Adam optimizer. The learning rate is set to 3×10^{-4} for the classifier and 3×10^{-5} for the ResNet-34 backbone.

The training is conducted in two phases: for the first 100 epochs, all parameters are trainable; in the final 30 epochs, only the classifier is fine-tuned. During inference, the final prediction is determined by majority voting across the three models.

4.3. Comparison with other entries

Table 1 summarizes the Top 3 results of the 3rd MiGA-IJCAI Challenge Track 3. Our method ranks third with an Accuracy of 63.46%, matching the score of the second-place method. Notably, while other leading methods may exploit additional modalities or fuse multi-source information, our approach relies solely on RGB inputs. Despite this, it achieves strong and competitive performance, highlighting the robustness and efficiency of our design, particularly in scenarios where access to extra modalities is limited or impractical.

Table 1
Comparison with other entries

Method	Accuracy (%)
backpacker	**69.23**
ISPCAST	63.46
haozhe bu (Ours)	63.46

4.4. Ablation Study

Our method ranked third in the IJCAI 2025 MiGA Challenge, demonstrating its strong capacity to tackle complex, real-world emotion recognition tasks. To analyze the contribution of each module, we conduct a comprehensive ablation study across three aspects: class balancing strategy, dual-channel feature modeling, and multi-model ensemble.

Comparison of Class Balancing Strategies

Table 2 shows the classification accuracies for three commonly used class imbalance handling strategies.

Table 2
Comparison of class balancing strategies

Method	Accuracy (%)
Baseline (No Balancing)	55.77
Oversampling	53.85
Loss Reweighting	56.73
Intra-Class Partitioning (Ours)	**61.54**

Given an approximate class imbalance between the “win" and “loss" categories, naive oversampling—which duplicates minority class samples to balance class counts—leads to a decline in accuracy to 53.85%. This suggests that oversampling may distort the original data distribution and introduce overfitting.

Loss reweighting slightly improves upon the baseline, reaching an accuracy of 56.73% by assigning greater gradient weight to the minority class during training. While this method provides some relief, its impact remains limited and indirect, particularly for deep models.

In contrast, the proposed intra-class partitioning strategy divides the majority class into two disjoint subsets and trains two identical yet independent models on the resulting balanced data. This method neither alters the data distribution nor relies on synthetic samples. It achieves the highest accuracy of 61.54%, demonstrating superior effectiveness and robustness in behavior-based emotion recognition for tennis match outcome prediction.

Effectiveness of Dual-Channel Image Feature Extraction

To evaluate the impact of dual-channel image features, we compare three input settings—scene only, face only, and combined scene+face—under both balanced and imbalanced training. Results are shown in Table 3.

Table 3
Comparison of Different Feature Inputs and Data Balance Settings

Setting	Scene	Face	Scene + Face
Balanced	51.92	54.81	**61.54**
Imbalanced	48.08	53.85	**55.77**

In the balanced setting, dual-channel input significantly outperforms both single-channel variants, reaching 61.54%. This validates our feature modeling strategy, where the ResNet34 backbone jointly encodes scene-level actions and facial micro-expressions. Their fusion effectively captures complementary behavioral signals.

Even under class imbalance, the dual-channel input remains more robust than either single modality, achieving 55.77%. These results highlight the synergy between proper class balancing and dual-view feature design.

It is worth noting that the scene-only input yields the lowest performance across both settings, likely due to sparse behavioral cues and background interference. The face-only model improves over scene-only, but still struggles when athletes deliberately suppress facial expressions. In contrast, combining face and scene provides a balanced and expressive representation, demonstrating superior reliability.

Performance of Multi-Model Ensemble Strategy

We further assess the contribution of ensemble learning by comparing the individual model accuracies with the ensemble result. Table 4 presents the results.

Table 4
Performance comparison of individual models and ensemble

Model	**Accuracy (%)**
BalanceNet A	61.54
BalanceNet B	60.58
EmotionScoreNet (DeepFace)	59.62
Ensemble	**63.46**

Both BalanceNet A and B are trained on different partitions of the majority class but share identical architectures. Their accuracy reflects consistency and diversity. The EmotionScoreNet, based on DeepFace, provides semantic-level priors and performs comparably.

When combined via majority voting, the ensemble model achieves 63.46% accuracy, surpassing all individual models. This demonstrates that ensemble learning successfully leverages structural diversity and complementary information, especially under ambiguous emotional signals.

5. Conclusion

This paper presents a behavior-based emotion recognition framework for class-imbalanced video data, proposed for the IJCAI 2025 MiGA Challenge. The goal is to perform emotion recognition and match outcome prediction in post-match interview scenarios. The method comprehensively addresses real-world challenges such as imbalanced data distribution, emotional masking, and weak behavioral signals, by constructing a complete framework that integrates intra-class partitioning, dual-channel image feature extraction, emotional prior modeling, and multi-model ensemble decision-making.

Specifically, the proposed intra-class partitioning strategy alleviates class imbalance without altering the original data distribution, effectively improving the model's capacity to learn from minority class samples. Meanwhile, the dual-channel feature extraction mechanism, based on full-body actions (scene) and facial expressions (face), integrates both global and local temporal cues, enhancing the model's sensitivity to complex behaviors and fine-grained emotional signals. In addition, an external emotional prior module based on DeepFace is incorporated to provide complementary decision cues to the visual pathway. Finally, a multi-model voting ensemble is used to improve overall robustness and generalization. Experimental results show that the proposed method achieves strong performance in the MiGA Challenge and exhibits significant advantages in all ablation studies, demonstrating its effectiveness and practicality in complex real-world tasks.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4 and Qwen in order to: grammar and spelling check. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

Acknowledgement

The work is jointly supported by the National Natural Science Foundation of China under grants No.62472342, and 62272364, the National Science and Technology Major Project under grant No.2022ZD0117103, the provincial Key Research and Development Program of Shaanxi under grant No.2024GH-ZDXM-47, the Research Project on HigherA Education Teaching Reform of Shaanxi Province under grant No.23JG003, the Fundamental Research Funds for the Central Universities under grant No.QTZX25037.

References

[1] Yunan Li et al. "Learning Robust Representations with Information Bottleneck and Memory Network for RGB-D-based Gesture Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).* Oct. 2023, pp. 20968–20978.

[2] Yunan Li et al. "Cr-net: A deep classification-regression network for multimodal apparent personality analysis". In: *International Journal of Computer Vision* 128 (2020), pp. 2763–2780.

[3] Xin Hu et al. "Modality Fusion Adaptor-Enhanced Vision Transformer for Multimodal Action Recognition". In: *International Conference on Pattern Recognition.* Springer. 2025, pp. 314–323.

[4] Sefik Serengil and Alper Ozpinar. "A Benchmark of Facial Recognition Pipelines and Co-Usability Performances of Modules". In: *Journal of Information Technologies* 17.2 (2024), pp. 95–107. DOI: 10.17671/gazibtd.1399077. URL: https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077.

[5] Sefik Ilkin Serengil and Alper Ozpinar. "LightFace: A Hybrid Deep Face Recognition Framework". In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU).* IEEE. 2020, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802. URL: https://ieeexplore.ieee.org/document/9259802.

[6] Sefik Ilkin Serengil and Alper Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *2021 International Conference on Engineering and Emerging Technologies (ICEET).* IEEE. 2021, pp. 1–4. DOI: 10.1109/ICEET53442.2021.9659697. URL: https://ieeexplore.ieee.org/document/9659697.

[7] Sefik Serengil and Alper Ozpinar. "Encrypted Vector Similarity Computations Using Partially Homomorphic Encryption: Applications and Performance Analysis". In: *arXiv preprint arXiv:2503.05850* (2025). doi: 10.48550/arXiv.2503.05850. [Online]. Available: https://arxiv.org/abs/2503.05850.

[8] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[9] Yin Cui et al. “Class-balanced loss based on effective number of samples”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9268–9277.

[10] Bingyi Kang et al. “Decoupling representation and classifier for long-tailed recognition”. In: *arXiv preprint arXiv:1910.09217* (2019).

[11] Shan Li, Weihong Deng, and JunPing Du. “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[12] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. “Cross-dataset learning and person-specific normalisation for automatic action unit detection”. In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 6. IEEE. 2015, pp. 1–6.

[13] Behzad Hasani and Mohammad H Mahoor. “Facial expression recognition using enhanced deep 3D convolutional neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2017, pp. 30–40.

[14] Ciprian Adrian Corneanu et al. “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications”. In: *IEEE transactions on pattern analysis and machine intelligence* 38.8 (2016), pp. 1548–1568.

[15] Xin Liu et al. “imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 10631–10642.

A. Online Resources

- Our code is available at the GitHub repository: https://github.com/wudidaluobo/MiGA-IJCAI