Multi-Track Multimodal Learning on iMiGUE: Micro-Gesture and Emotion Recognition

Arman Martirosyan[1,*], Shahane Tigranyan[1], Maria Razzhivina[2], Artak Aslanyan[5], Nazgul Salikhova[6], Ilya Makarov[2,3], Andrey Savchenko[2,4] and Aram Avetisyan[2,7]

[1]Russian - Armenian University, Yerevan, Armenia
[2]ISP RAS, Moscow, Russia
[3]AIRI, Moscow, Russia
[4]Sber AI Lab, Moscow, Russia
[5]HSE University, Moscow, Russia
[6]Innopolis University, Innopolis, Russia
[7]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia

Abstract

Micro-gesture recognition and behavior-based emotion prediction are both highly challenging tasks that require modeling subtle, fine-grained human behaviors, primarily leveraging video and skeletal pose data. In this work, we present two multimodal frameworks designed to tackle both problems on the iMiGUE dataset. For micro-gesture classification, we explore the complementary strengths of RGB and 3D pose-based representations to capture nuanced spatio-temporal patterns. To comprehensively represent gestures, video, and skeletal embeddings are extracted using MViTv2-S and 2s-AGCN, respectively. Then, they are integrated through a Cross-Modal Token Fusion module to combine spatial and pose information. For emotion recognition, our framework extends to behavior-based emotion prediction, a binary classification task identifying emotional states based on visual cues. We leverage facial and contextual embeddings extracted using SwinFace and MViTv2-S models and fuse them through an InterFusion module designed to capture emotional expressions and body gestures. Experiments conducted on the iMiGUE dataset, within the scope of the MiGA 2025 Challenge, demonstrate the robust performance and accuracy of our method in the behavior-based emotion prediction task, where our approach secured 2nd place.

Keywords

Behavior-based emotion recognition, micro-gesture, action classification, video understanding, multimodal learning

1. Introduction

Micro-gesture recognition and behavior-based emotion prediction are both challenging and high-impact tasks that aim to interpret fine-grained human behaviors from visual data. These tasks are foundational for next-generation applications in human-computer interaction, affective computing, sign language interpretation, and immersive virtual or augmented reality environments. Despite their shared reliance on subtle behavioral cues, they differ in scope and modality requirements: micro-gesture recognition emphasizes the detection of low-amplitude movements in fingers, hands, or facial muscles using both RGB video and skeletal pose data, while behavior-based emotion prediction focuses solely on facial and contextual cues from video to infer emotional states in real-world settings, such as post-match interviews.

MiGA@IJCAI25: International IJCAI Workshop on 3rd Human Behavior Analysis for Emotion Understanding, August 29, 2025, Guangzhou, China.
*Corresponding author.
martirosyan.arman@student.rau.am (A. Martirosyan); shahane.tigranyan@rau.am (S. Tigranyan); mvrazzhivina@edu.hse.ru (M. Razzhivina); aaaslanyan_2@edu.hse.ru (A. Aslanyan); n.salikhova@innopolis.university (N. Salikhova); iamakarov@hse.ru (I. Makarov); avsavchenko@hse.ru (A. Savchenko); a.a.avetisyan@ispras.ru (A. Avetisyan)
0009-0002-0969-585X (A. Martirosyan); 0000-0003-1536-9954 (S. Tigranyan); 0000-0003-0188-6846 (M. Razzhivina); 0009-0004-8290-2557 (A. Aslanyan); 0009-0004-7246-280X (N. Salikhova); 0000-0002-3308-8825 (I. Makarov); 0000-0001-6196-0564 (A. Savchenko); 0000-0002-7066-6954 (A. Avetisyan)

Unlike coarse gestures or full-body actions, micro-gestures involve nuanced and often ambiguous motions that demand high temporal and spatial resolution for reliable classification. Similarly, in emotion prediction, the ability to infer affective states from visual cues without relying on speech or textual data requires a fine-grained understanding of temporally distributed behaviors and subtle facial dynamics. In both tasks, conventional single-modality approaches fall short due to the complexity and ambiguity of the underlying signals.

Recent works[1, 2, 3] have demonstrated that combining multiple modalities, such as RGB video and pose or skeleton-based representations, can improve robustness and discrimination in gesture recognition. However, most existing fusion methods based on simple concatenation[4], late fusion[5], or shallow attention[6] fail to fully exploit the fine-grained complementary information between modalities, especially when the differences between gesture classes are subtle and context-dependent.

To address these challenges, we integrate a novel framework that combines multi-modal token-level feature learning with context-aware class refinement for precise micro-gesture recognition. We leverage MViTv2-S[7] and 2s-AGCN[8] encoders to extract temporally-aware visual tokens from RGB video frames and 3D skeletal pose sequences, capturing modality-specific dynamics in high resolution. To unify these heterogeneous modalities, we employ a Cross-Modal Token Fusion module that aligns and merges cross-modal tokens using multiple fusion heads based on spatial, semantic, and contextual relevance.

Furthermore, to refine class decision boundaries, we incorporate a Memory-Powered Refinement Module that learns to refine gesture classification based on accumulated knowledge of gesture representations.

We evaluate our approach on the iMiGUE[9] dataset and observe strong performance, demonstrating its effectiveness for fine-grained micro-gesture classification.

In addition to micro-gesture recognition, behavior-based emotion classification represents a complementary and equally challenging task in affective computing. Unlike conventional emotion recognition approaches that rely on facial expressions or vocal cues, behavior-based emotion prediction seeks to infer an individual's emotional state based on nonverbal visual signals such as body posture and micro-gestures. Accurately modeling such subtle, temporally distributed behavioral patterns requires not only robust feature extraction from multiple visual streams but also effective fusion strategies that can capture inter-modal dependencies. In this work, we extend our framework to address this task using a dual-stream architecture that combines facial and contextual embeddings through iterative gated fusion.

To evaluate the effectiveness of our approach in this setting, we conducted extensive experiments on the iMiGUE dataset, which provides a suitable benchmark for understanding behavior-driven emotions in real-world scenarios.

2. Micro-gesture Classification

2.1. Task Definition

The Micro-Gesture Classification task is a 32-way classification problem defined on short video clips containing fine-grained, subtle hand gestures. Given a video segment V with T frames, the goal is to predict a gesture label $y \in \{0, 1, \ldots, 31\}$, where each class corresponds to a distinct micro-gesture, and class 99 indicates a *non-gesture* (i.e., non-illustrative gesture).

2.2. Model Architecture

We propose a novel multimodal framework (Figure 1a) for fine-grained micro-gesture classification, which leverages both RGB and 3D skeletal pose representations. Our method builds upon recent advances in token-level fusion and multimodal refinement, integrating ideas from Multi-Criteria Token Fusion (MCTF)[10] and Context-Aware Prompt Learning (CAPL) [11] to improve alignment and discriminability of cross-modal features.

2.3. Modality Encoding

We extract modality-specific features using MViTv2-S and 2s-AGCN backbones. The RGB stream processes spatio-temporal clips of raw gesture videos, capturing fine-grained motion and appearance cues. The skeleton stream operates directly on 3D joint coordinates, using features extracted from a pretrained 2s-AGCN model.

2.4. Cross-Modal Token Fusion

To effectively align and integrate the RGB and skeleton features, we apply the Cross-Modal Token Fusion module (CMTF) inspired by [10]. The module performs token-level cross-modal attention by considering multiple semantic and spatial criteria, dynamically attending to the most relevant tokens from the complementary modality.

Given token sequences $\mathbf{T}_{\text{RGB}}$ and $\mathbf{T}_{\text{Pose}}$ extracted from the RGB and skeleton MViTv2 branches respectively, the Cross-Modal Token Fusion (CMTF) module produces a fused representation $\mathbf{T}_{\text{fused}}$ through dynamic token alignment:

$$\mathbf{T}_{\text{fused}} = \text{CMTF}(\mathbf{T}_{\text{RGB}}, \mathbf{T}_{\text{Pose}})$$

which is passed through a linear projection and temporal pooling to yield a compact feature vector for each gesture clip.

2.5. Memory-Powered Refinement Module

To further improve class separability, especially in fine-grained microgesture scenarios, we incorporate a Memory-Powered Refinement Module. This module maintains an external memory bank of prototypical embeddings per class. Inspired by prototype refinement approaches, it uses memory to refine predictions by comparing incoming features against stored high-confidence class exemplars.

During the initial training epochs, the memory is populated with confident feature embeddings. In later epochs, for each input, we compare its features with the top-k similar memory vectors (per predicted class) using cosine similarity. The comparisons are processed via multi-head self-attention to refine the features before classification. A refinement loss (L_p) is then computed, encouraging alignment with the most representative class features, and is combined with the standard classification loss (L_c):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{c}} + \alpha \mathcal{L}_{\text{p}} \tag{1}$$

Here, $\mathcal{L}_{\text{c}}$ supervises classification based on averaged modality logits, while $\mathcal{L}_{\text{p}}$ encourages tighter intra-class clustering and increased inter-class margins in the feature space through both parametric (prototypes) and non-parametric (external memory) constraints.

To balance modality contributions in the final decision, we adopt a weighted late fusion strategy. While both RGB and pose classifiers output independent logits, we place higher emphasis on the pose branch due to its robustness in capturing subtle skeletal dynamics in microgestures. The final class prediction is computed as:

$$\mathbf{C}_i = w_{\text{pose}} \cdot \mathbf{y}_{\text{pose}} + w_{\text{RGB}} \cdot \mathbf{y}_{\text{RGB}}, \quad \text{where} \quad w_{\text{pose}} > w_{\text{RGB}} \tag{2}$$

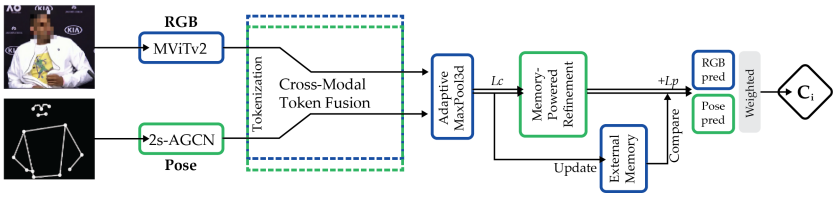
These weights are adjusted dynamically during training.

3. Behavior-based Emotion Prediction

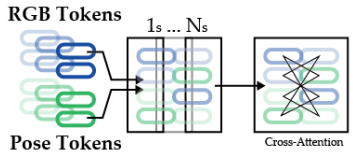
3.1. Task Definition

The Behavior-Based Emotion Recognition task[1] is a binary classification problem defined on video sequences of post-match interviews. Given an interview video clip V containing T frames, the goal is to

[1] https://www.kaggle.com/competitions/the-3rd-mi-ga-ijcai-challenge-track-3



(a) Overall architecture diagram.



(b) Cross-Modal token fusion module.

Figure 1: Detailed architecture diagrams for microgesture classification model.

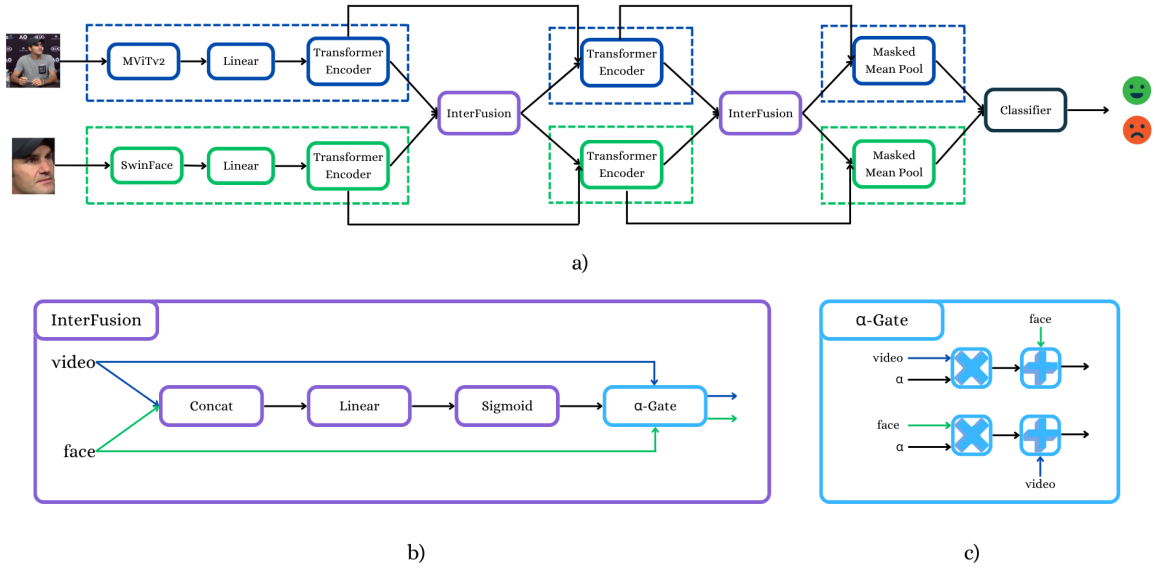


Figure 2: (a) Architecture of the proposed multimodal architecture for emotion recognition from video and facial features. (b) The structure of the InterFusion module. (c) The α-Gate mechanism for information aggregation from two modalities.

predict the match outcome label $y \in \{0, 1\}$, where $y = 1$ indicates a win and $y = 0$ - a loss. Each video contains visible body and facial behaviors that may implicitly express the emotional state of the athlete. The data is drawn from the iMiGUE dataset (details in Section), which captures fine-grained behavioral cues in press conferences of tennis players. The task is to develop a model that infers emotional signals relevant to the final outcome using only visual information, without access to audio or transcripts.

3.2. Model Architecture

We propose a dual-stream transformer-based architecture for behavior-based emotion recognition that integrates contextual and facial information in parallel and integrates them through interfusion blocks. The model is designed to capture both intra-modal dynamics and inter-modal interactions through fusion mechanisms. The architecture is shown in Figure 2(a).

The model takes as input two sequences:

- **Contextual embeddings:** The video is processed using a pretrained MViTv2-S backbone, which operates on non-overlapping chunks of 16 frames and produces one embedding per chunk. This results in a sequence of frame-group representations $\mathbf{X}^{\text{CTX}} \in \mathbb{R}^{T \times 768}$, where T denotes the number of 16-frame segments extracted from the video.
- **Facial embeddings :** Faces are detected in each frame using a pretrained YOLO-face model[2],

[2]https://github.com/akanametov/yolo-face

and cropped face regions are encoded using SwinFace[12], a hierarchical vision transformer tailored for facial representation learning. While SwinFace produces one embedding per frame, we aggregate the features by averaging every 16 consecutive frame-level embeddings, resulting in a sequence $\mathbf{X}^{\text{FACE}} \in \mathbb{R}^{T \times 512}$ that is temporally aligned with $\mathbf{X}^{\text{CTX}}$. This ensures one embedding per 16-frame window for both modalities, enabling consistent cross-modal fusion.

Each stream begins with a linear projection followed by dropout for regularization. The projected embeddings are then passed through a modality-specific Transformer Encoder to get intra-modal temporal dependencies.

To facilitate interaction between modalities, we employ the InterFusion module Figure 2(b), a lightweight gated fusion block that enables bidirectional information exchange between the contextual and facial streams. Rather than using attention, InterFusion computes a shared element-wise gate from the concatenated inputs and performs a residual gated-sum fusion in both directions (shown in Figure 2(c)). This allows each stream to selectively incorporate features from the other while preserving its own representation. This fusion step is followed by residual addition to preserve the original intra-modal information. The fused features are processed by an additional Transformer Encoder layer per stream, followed by a second InterFusion module and another residual connection. This iterative process allows deeper and progressively refined integration between the two modalities.

After temporal modeling, a masked mean pooling operation is applied across time for both streams, producing fixed-size representations. These are concatenated and passed through a classification head to predict the class.

This design enables the model to effectively capture subtle, temporally distributed cues from both body posture and facial expressions, which are critical in emotion-based outcome prediction.

4. Experiments

5. Dataset: iMiGUE

To evaluate our proposed methods, we utilized the iMiGUE (Identity-free Micro-Gesture Understanding and Emotion) dataset, introduced by Liu et al.[9]. iMiGUE is a large-scale, identity-free video dataset specifically designed to facilitate research in micro-gesture recognition and emotion analysis.

The dataset comprises 359 post-match press conference videos, featuring 72 athletes. Each clip captures spontaneous upper-body micro-gestures, such as "covering the face," "folding arms," or "crossing fingers," which are indicative of the athlete's emotional state following a win or loss.

Annotations in iMiGUE encompass 32 behavioral classes, including 31 distinct micro-gesture categories and one non-micro-gesture category. In addition to gesture labeling, each sample is also annotated with a binary emotion label (positive or negative) based on contextual interpretation (e.g., winning or losing a match) and observable behavior. Each video frame is processed using the OpenPose toolkit to extract skeletal data. This skeletal representation includes two-dimensional spatial coordinates and confidence scores for each joint, facilitating detailed analysis of body movements.

The dataset is split into predefined training, validation, and test sets to ensure consistent evaluation. Importantly, the test set is identity-independent and includes subjects that are not present in the training and validation sets. This split emphasizes generalization and mitigates overfitting to specific individuals, thus supporting the development of identity-agnostic models for micro-gesture and emotion recognition. For the emotion classification task, the test set comprises 100 videos, divided between 50 positive and 50 negative emotional states.

To address the challenge of imbalanced class distributions inherent in spontaneous behavior datasets, the authors propose an unsupervised learning method that aims to capture latent representations from the micro-gesture sequences themselves. This approach enhances the model's ability to generalize across diverse gesture categories and improves emotion recognition performance.

In the context of the MiGA 2025 Challenge, the organizers released an enhanced version of the iMiGUE dataset that includes unblurred facial regions. This version aims to facilitate comprehensive

multimodal analysis by allowing researchers to incorporate facial expressions alongside body gestures in emotion recognition tasks. The availability of unblurred faces enables the development and evaluation of models that leverage both facial and gestural cues, potentially leading to more accurate and robust emotion understanding systems.

5.1. Microgesture Classification Experiments

Our experiments were conducted using a custom transformer-based classifier that integrates a Cross-Modal Token Fusion module and a refinement module with external memory. The model takes RGB and skeleton features from pre-trained MViTv2 and 2s-AGCN encoders and fuses them using multi-head self-attention to capture token-level cross-modal interactions. The fused representations are passed through adaptive pooling and modality-specific classifiers, with logits combined using a weighted sum for the final prediction.

To refine class boundaries, we introduce a prototypical memory module trained with a two-stage loss. During early epochs, the prototypical loss is disabled (α=0.0) to allow memory buildup. Its influence is then gradually activated (α=1.0) to enhance class separation via learned prototypes.

The final model configuration hyperparameters include: a hidden size of 512, 8 transformer heads, a memory size of 50 entries per class with top-5 nearest prototypes used for refinement, and a momentum of 0.9 for memory updates. Training was performed using the AdamW optimizer with a learning rate of 1e-4, weight decay of 1e-4, and ReduceLROnPlateau scheduler.

To ensure stable training and better generalization, variable-length sequences were padded dynamically per batch; bucketing and balanced sampling were used to mitigate class imbalance.

Table 1
Micro-gesture classification results on the iMiGUE.

Method	Modality	Top-1 Accuracy (%)
TSM [13]	RGB	58.77
VSwim-T [14]	RGB	59.97
VSwim-S [14]	RGB	57.83
VSwim-B [14]	RGB	61.73
ST-GCN [15]	Skeleton	46.38
ST-GCN++ [16]	Skeleton	49.56
StrongAug [16]	Skeleton	53.13
AAGCN [17]	Skeleton	54.73
CTR-GCN [18]	Skeleton	53.02
DG-STGCN [19]	Skeleton	49.56
PoseConv3D [20]	Skeleton	61.11
DSCNet [21]	RGB & Skeleton	62.53
Ours	RGB & Skeleton	**62.87**

Table 2
Fusion method and MR* impact on accuracy.

Fusion Method	Top-1 Accuracy (%)
Late Fusion	58.81
CMTF* (w/o MR*)	62.23
CMTF* (w/ MR*)	**62.87**

CMTF*: Cross-Modal Token Fusion;
MR*: Memory Refinement Block.

5.2. Emotion Prediction Experiments

For the Track 3 task of the MiGA 2025 Challenge, the proposed multimodal fusion model was evaluated on the enhanced iMiGUE dataset, which includes unblurred facial regions.

To enhance model performance, a comprehensive hyperparameter optimization was performed over the following search space: learning rate (1e-5 to 5e-4), transformer encoder depth (1 to 8), number of attention heads (2 to 8), dropout rate (0.1 to 0.5), and Focal Loss γ parameter (0.5 to 1.5). To address a severe class imbalance in the validation set, which initially contained only positive emotion samples, five negative samples were manually transferred from the training set to the validation set and excluded from training. This adjustment ensured more reliable validation performance across both classes. Class weights were calculated inversely proportional to the class frequencies in the adjusted training set.

Both Binary Cross-Entropy (BCE) and Focal Loss were evaluated, with Focal Loss demonstrating superior performance under the imbalanced setting.

The final model configuration used for comparison included a hidden size of 512, transformer encoder depth of 8, 4 attention heads, and a dropout rate of 0.5. The Focal Loss γ parameter was set to 0.5.

Training was conducted for up to 20 epochs with a batch size of 8, a learning rate of 1e-7, and early stopping with a patience of 7 epochs.

All experiments were conducted on a single NVIDIA A100 GPU with 80GB VRAM. Table 3 summarizes the comparison between our method, baseline approaches, and other submissions.

Table 3
Leaderboard results from MiGA Challenge Track 3.

Rank	Team	Score
1	backpacker	0.69230
2	**ISPCAST** (ours)	0.63461
3	haozhe bu	0.63461
4	gkdx2	0.62500
5	KeXu2233	0.60576
6	Baseline	0.39423

6. Conclusion

In this work, we presented two multimodal learning frameworks for addressing micro-gesture classification and behavior-based emotion recognition on the iMiGUE dataset. For the micro-gesture task, we proposed a novel architecture that fuses RGB and skeleton modalities via Cross-Modal Token Fusion and refines predictions through a memory-powered module leveraging class prototypes. Our method demonstrated strong performance, outperforming several prior RGB and skeleton-based baselines.

In the behavior-based emotion recognition task, we presented a dual-stream transformer-based model that jointly leverages facial and contextual visual information. The architecture incorporates iterative gated fusion via InterFusion modules to enable deep cross-modal interaction while preserving intra-modal information. The proposed model demonstrated competitive results in the MiGA 2025 Challenge Track, securing second place on the official leaderboard.

These results highlight the importance of fine-grained spatio-temporal modeling and multimodal interaction for subtle behavior understanding.

Acknowledgment

This work was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: grammar, spelling check, and paraphrase. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] H. Liu, X. Zhang, M. Yu, et al., Prototype learning for micro-gesture classification, arXiv preprint arXiv:2408.03097 (2024).

[2] K. Chen, X. Li, J. Wang, Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers, arXiv preprint arXiv:2403.10030 (2024).

[3] C.-F. Chen, Q. Xie, M.-H. Niu, et al., Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.
[4] B. Seddik, S. Gazzah, N. E. B. Amara, Hybrid multi-modal fusion for human action recognition, in: ICIAR, 2017.
[5] Y. Liu, Q. Chen, Late fusion for robust gesture recognition, in: ICPR, 2020.
[6] Y. Tan, B. Wang, Shallow attention mechanisms in multimodal learning, in: Proceedings of the 2021 International Conference on Multimodal Interaction, 2021, pp. 45–52.
[7] Y. e. a. Li, Mvitv2: Improved multiscale vision transformers for classification and detection, in: CVPR, 2022.
[8] L. Li, X. Zhang, Y. Song, Y. Chen, J. Wang, T. Yang, R. Lee, J. Wang, J. Huang, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12026–12035.
[9] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, G. Zhao, 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 (CVPR), 2021, pp. 13389–13398.
[10] S. Lee, J. Choi, H. J. Kim, Multi-criteria token fusion with one-step-ahead attention for efficient vision transformers, in: Conference on Computer Vision and Pattern Recognition, 2024.
[11] X. Lin, Y. Zhou, Y. Wang, Q. Huang, G. Huang, Context-aware prompt learning for vision-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
[12] L. Qin, M. Wang, C. Deng, K. Wang, X. Chen, J. Hu, W. Deng, Swinface: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation, IEEE Transactions on Circuits and Systems for Video Technology 34 (2023) 2223–2234.
[13] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7083–7093.
[14] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 3202–3211.
[15] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
[16] H. Duan, J. Wang, K. Chen, D. Lin, Pyskl: Towards good practices for skeleton action recognition, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 7351–7354.
[17] L. Shi, Y. Zhang, J. Cheng, H. Lu, Skeleton-based action recognition with multi-stream adaptive graph convolutional networks, IEEE Transactions on Image Processing 29 (2020) 9532–9545.
[18] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 13359–13368.
[19] H. Duan, J. Wang, K. Chen, D. Lin, Dg-stgcn: Dynamic spatial-temporal modeling for skeleton-based action recognition, arXiv preprint arXiv:2210.05895 (2022).
[20] H. Duan, Y. Zhao, K. Chen, D. Lin, B. Dai, Revisiting skeleton-based action recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 2969–2978.
[21] Q. Cheng, J. Cheng, Z. Liu, Z. Ren, J. Liu, A dense-sparse complementary network for human action recognition based on rgb and skeleton modalities, Expert Systems with Applications 244 (2024) 123061. URL: https://www.sciencedirect.com/science/article/pii/S0957417423035637. doi:10.1016/j.eswa.2023.123061.