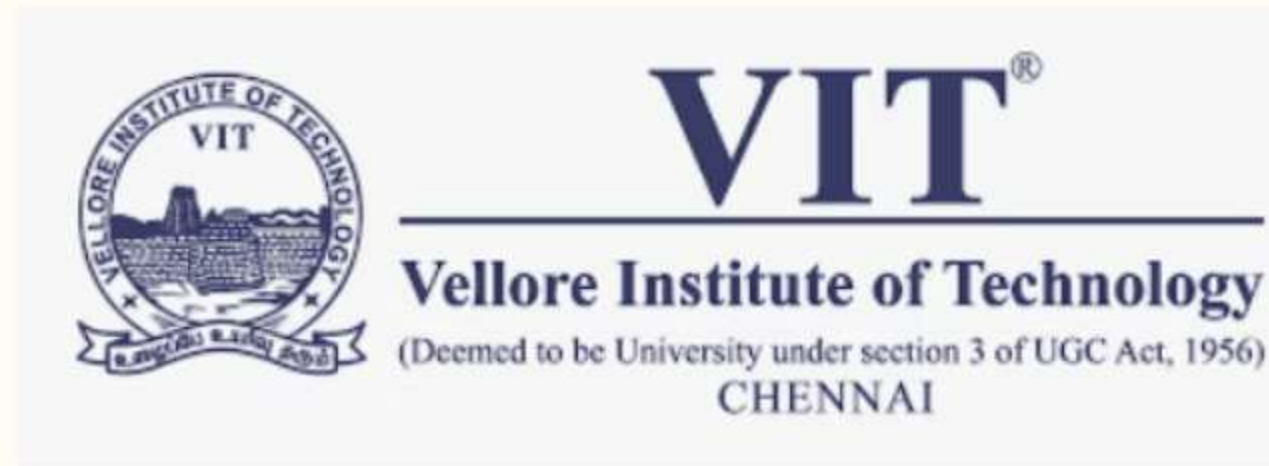


Vellore Institute Of Technology



DIGITAL IMAGE PROCESSING (BCSE403L)


Submitted to Professor. Dr. Geetha S

AYUSHI JHA (22BCE1980)

SOHAM AMBERKAR (22BCE1770)

Predicting Engagement and Emotional States in Educational Videos Using Temporal Facial Feature Analysis



 colab.research.google.com

Google Colab



This presentation covers our project on recognizing student engagement from facial expressions using a novel self-supervised learning method called Facial Masked Autoencoder (FMAE). This approach addresses challenges in online education by analyzing unlabeled facial video data to improve engagement detection accuracy and reduce annotation costs.

We explore the architecture, facial masking strategy, reconstruction module, and experimental results on benchmark datasets, demonstrating the effectiveness of FMAE in enhancing online learning experiences.

Introduction to Student Engagement Recognition



Importance

Student engagement is key to measuring learning effectiveness, especially in online education where real-time feedback is limited.



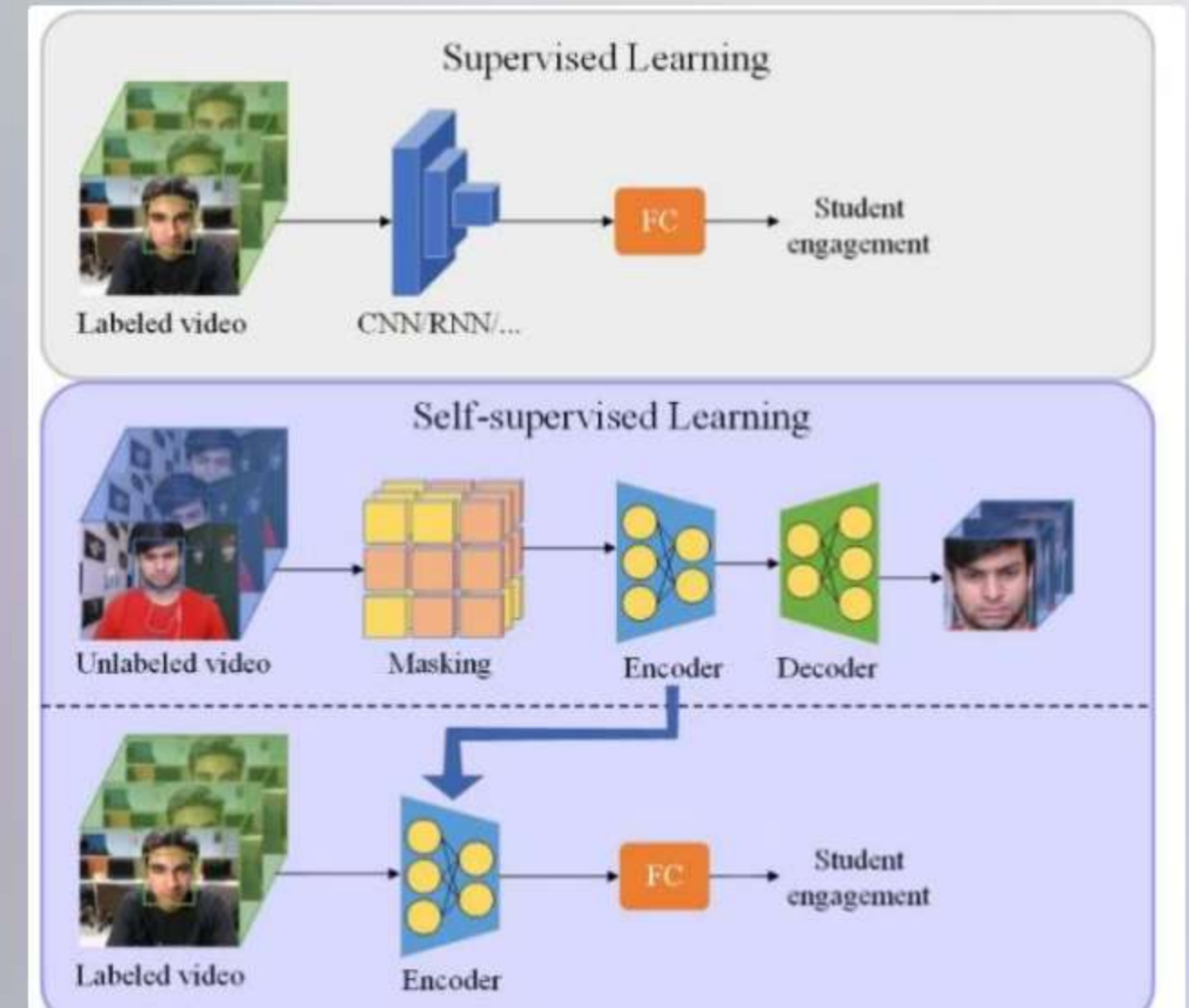
Challenges

Supervised methods require large labeled datasets, which are scarce and costly to produce, limiting model generalization.



Our Solution

We propose a self-supervised learning method that leverages unlabeled facial videos to learn rich facial features for engagement recognition.



Related Work: Trends in Engagement Recognition

3D CNNs

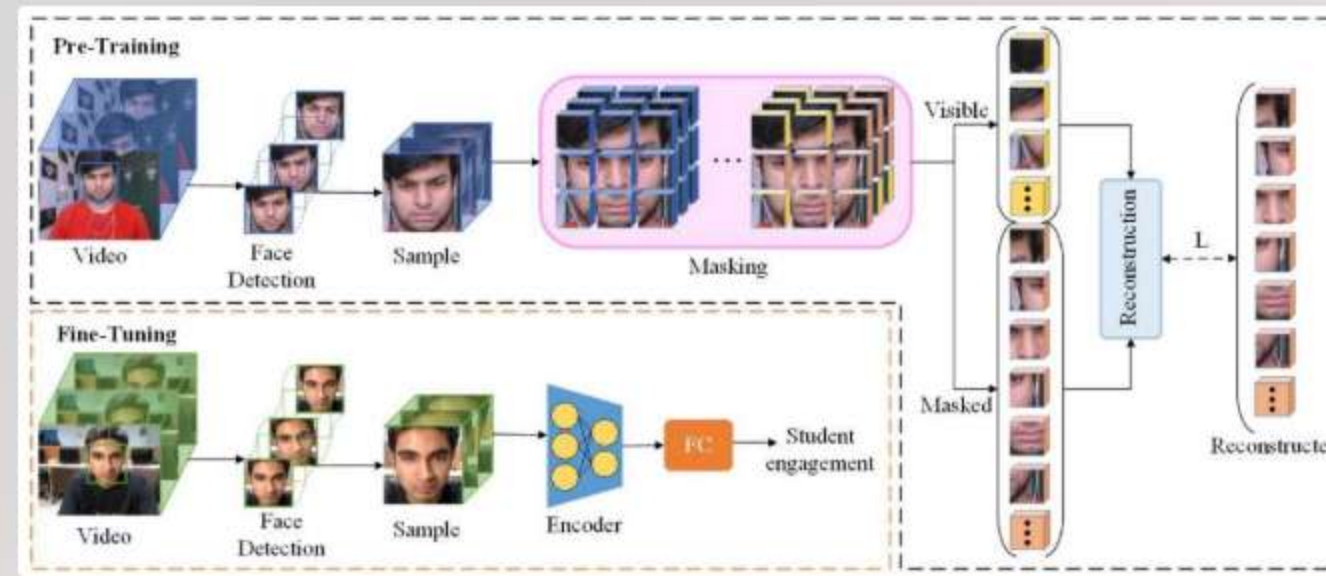
Extract spatio-temporal features directly from video frames for engagement detection.

Ensemble Models

Combine 2D CNNs for spatial features with RNNs for temporal dynamics to improve accuracy.

Transformer Architectures

Use global attention mechanisms to capture complex facial feature relationships over time.



Proposed Method: Facial Masked Autoencoder (FMAE)

1

Self-Supervised Pre-training

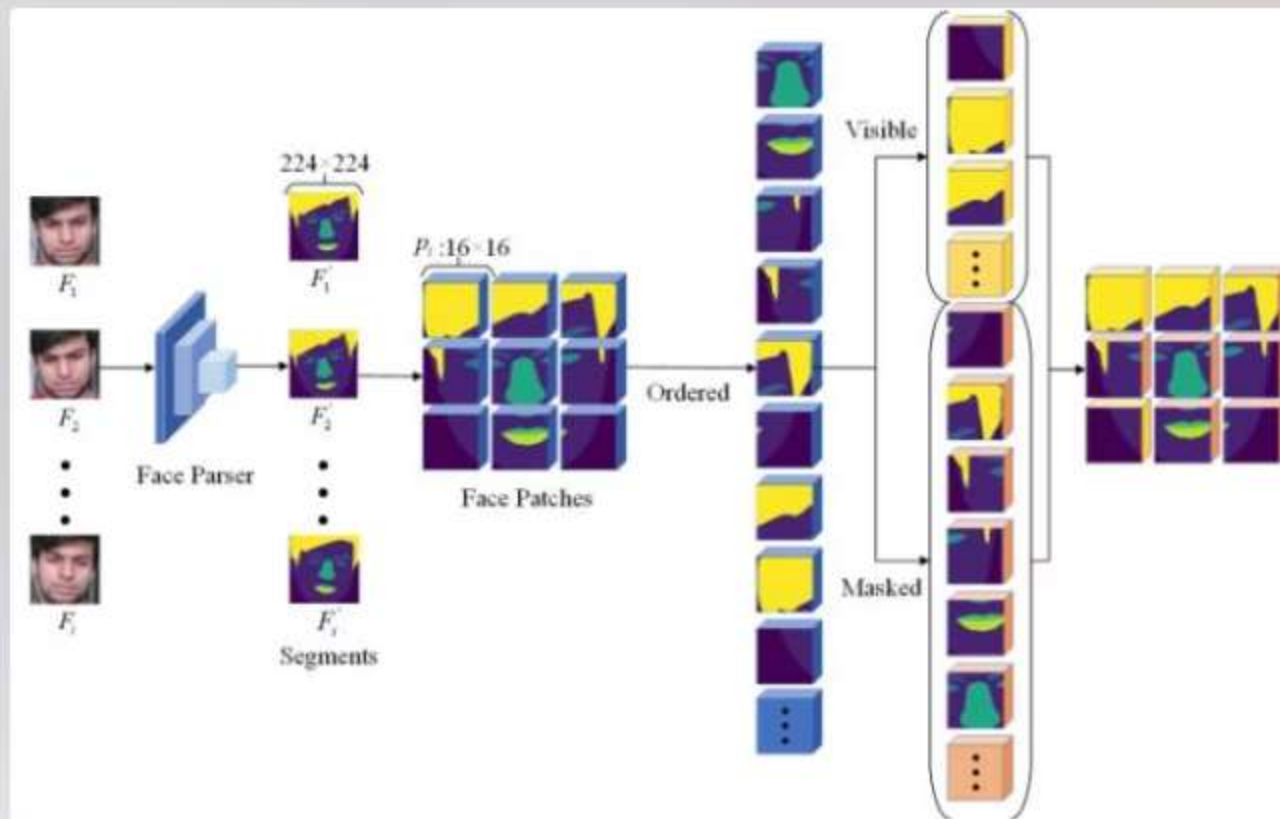
FMAE learns facial representations from unlabeled video data by reconstructing masked facial regions.

2

Fine-Tuning

The pre-trained encoder is fine-tuned with labeled data for student engagement recognition.

Facial Masking Strategy



Dynamic Spatio-Temporal Masking

Mask the same facial region consistently across time frames to prevent information leakage.

Region Prioritization

Higher masking priority is given to key facial areas like eyes, nose, and mouth for richer feature learning.

Selective Masking

Secondary regions such as hair and skin are masked later to focus reconstruction on important facial features.

Reconstruction Module



Encoder-Decoder Architecture

The encoder maps visible tokens to latent space; the decoder reconstructs masked facial regions.



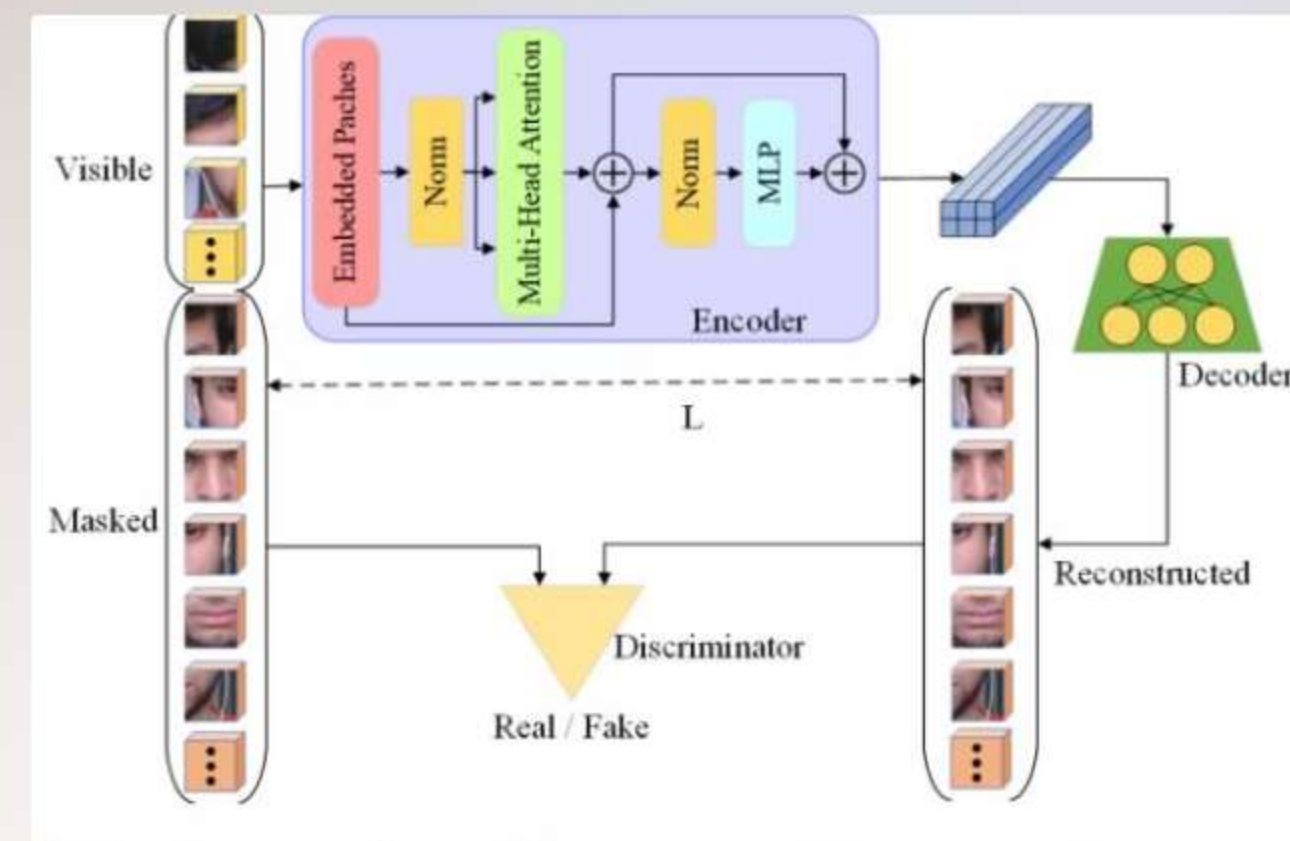
Training Process

Iterative training updates encoder, decoder, and discriminator to optimize facial feature learning.



Loss Functions

Combines reconstruction loss (pixel accuracy) and adversarial loss (realism) for clearer results.



Loss Functions Explained

Reconstruction Loss

Minimizes pixel-wise difference between original and reconstructed masked regions to ensure accuracy.

Adversarial Loss

Uses a discriminator to encourage the model to produce realistic and clear facial reconstructions.

Advantages of FMMAE for Online Learning



Reduced Annotation Cost

Leverages unlabeled data, minimizing the need for expensive manual labeling.



Focused Feature Learning

Facial masking strategy enhances learning of important facial regions and temporal details.



Improved Generalization

Self-supervised pre-training helps overcome overfitting and boosts model robustness.

Experimental Settings and Datasets



DAiSEE Dataset

Contains 9,068 videos from 112 online learners, labeled for boredom, confusion, frustration, and engagement at four levels. Videos are 10 seconds long at 30 fps, 640×480 resolution. The test set includes 1,784 videos.

EmotiW Dataset

Includes 262 videos from 78 people during online learning, with engagement levels from disengagement to high engagement. Videos last about 5 minutes at 30 fps, 640×480 resolution. Training and validation sets are publicly available.



Evaluation Metrics and Implementation Details

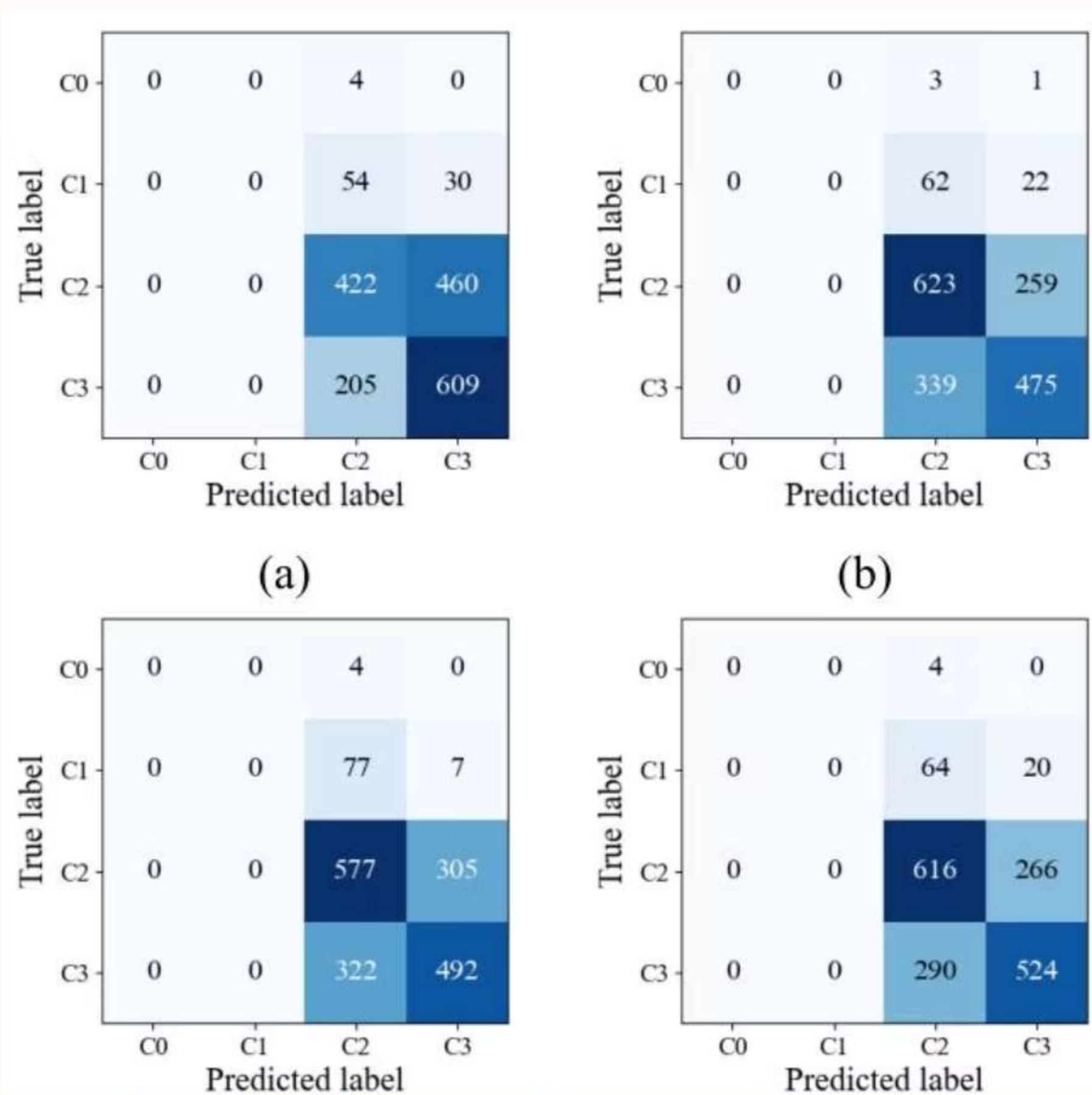
Metrics

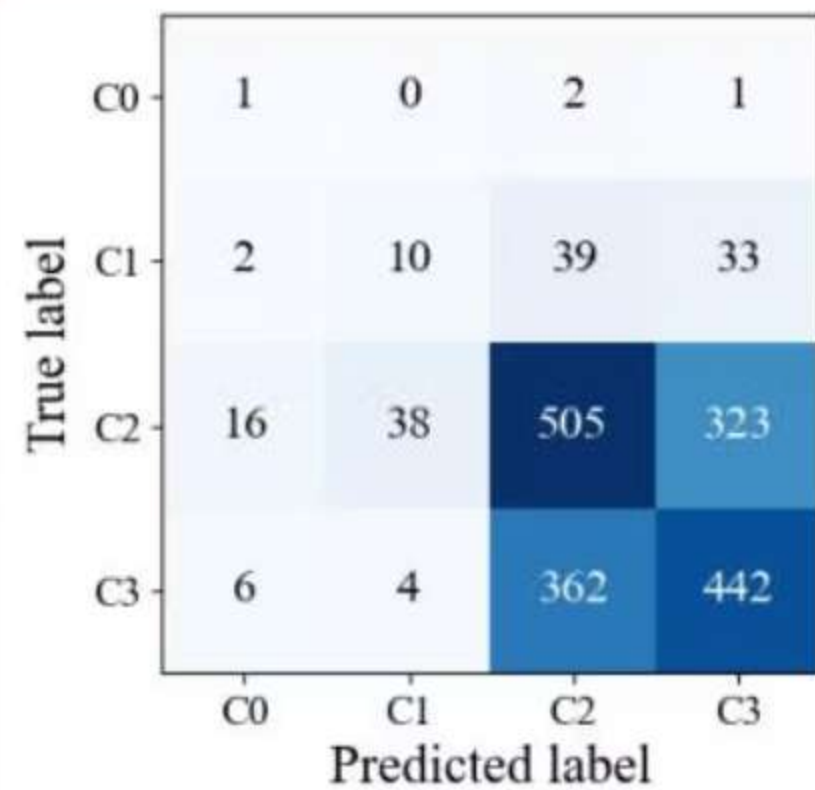
Accuracy, MSE, Precision, Recall, and F1-Score are used to evaluate model performance on engagement recognition tasks.

Implementation

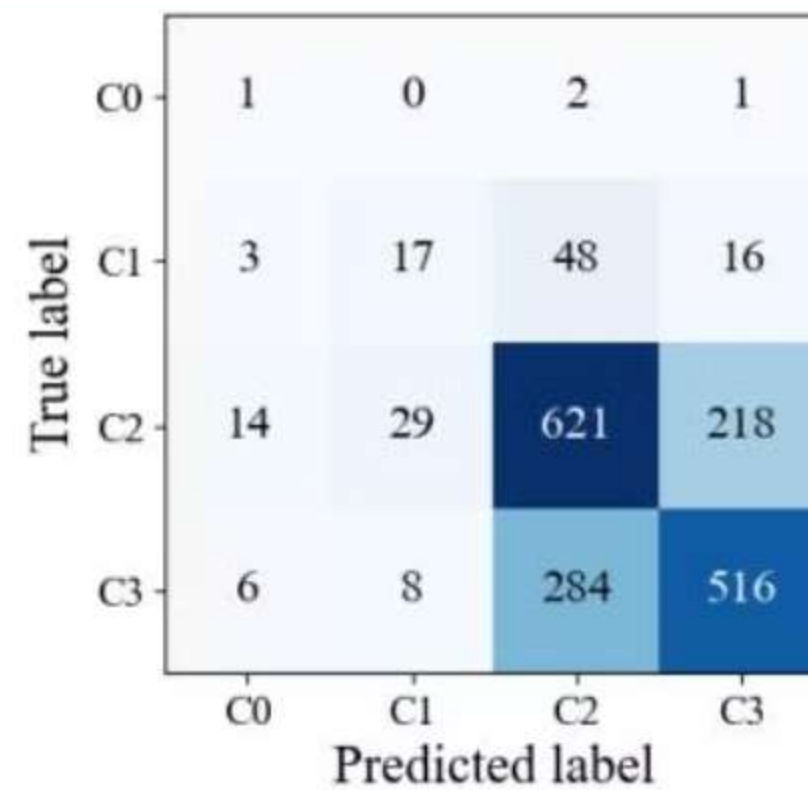
Videos are sampled with a temporal stride of 2 to reduce redundancy. FMAE masks 90% of spatio-temporal facial tokens, encoding them with ViT-B and reconstructing masked regions using combined reconstruction and adversarial losses.

Performance Comparison with State-of-the-Art





(e)



(f)

Dataset

FMAE achieves 64.74% accuracy, outperforming previous best methods like ResNet + TCN (63.90%) and Optimized ShuffleNet v2 (63.90%).

EmotiW Dataset

FMAE attains a competitive MSE of 0.0629, close to the best method MAGRU (0.0517), demonstrating strong facial representation learning for engagement recognition.

```

▶ # Apply attention and predict engagement:

attended_frames = apply_attention(all_frames)

if len(attended_frames) >= sliding_window_size:
    _, scores = fer.predict_engagement(attended_frames, sliding_window_size)
    score = np.mean(scores, axis=0)
    engagement_idx = np.argmax(score)

    engagement_levels = ["Poor", "Slight", "Moderate", "High", "Very High"]
    engagement_level_idx = int(round(engagement_idx * (len(engagement_levels) - 1)))
    engagement_level = engagement_levels[engagement_level_idx]

    confidence_percentage = (score[engagement_idx] / np.sum(score)) * 100

    print(f"Predicted Engagement Level: {engagement_level}")
    print(f"Confidence Percentage: {confidence_percentage:.2f}%")

else:
    print(f"Not enough frames to predict engagement. Required: {sliding_window_size}, Found: {len(attended_frames)}")

```

```

↵ Predicted Engagement Level: Very High
Confidence Percentage: 95.01%

```



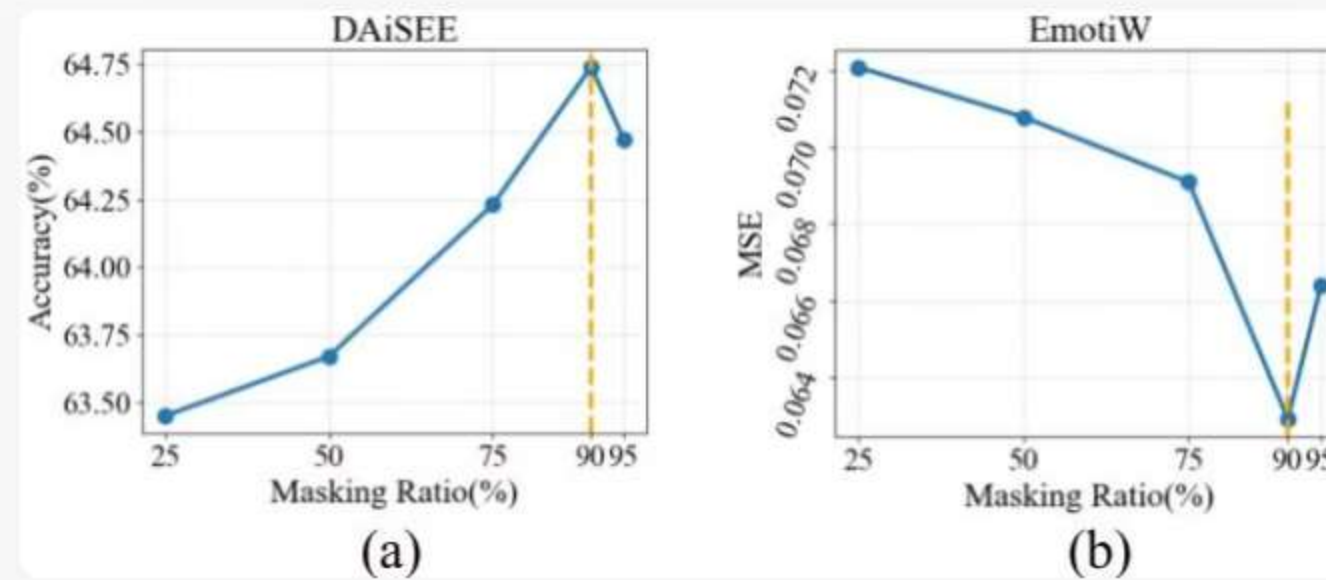
Ablation Study: Masking Ratio and Temporal Stride

Masking Ratio

Experiments show 90% masking ratio is optimal, balancing feature understanding and reconstruction difficulty. Lower or higher ratios reduce performance.

Minimum Temporal Stride

Stride of 2 offers a good trade-off, slightly reducing accuracy but significantly lowering computational cost, making it the most efficient choice.



Ablation Study: Masking Strategies and Modules

Masking Strategies

- Facial masking outperforms random, frame, and tube masking in accuracy and F1-score.
- Facial masking focuses on key facial regions, improving feature learning.

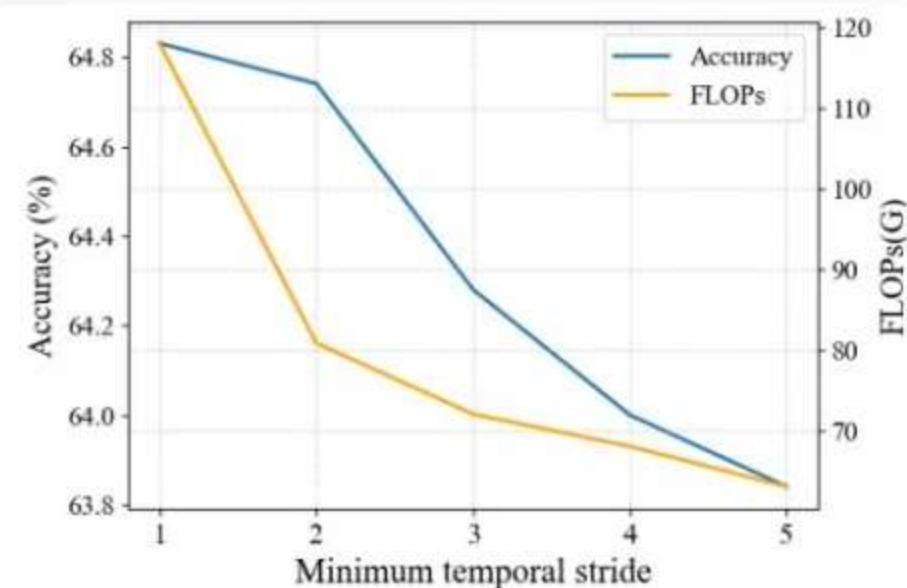
Module Effects

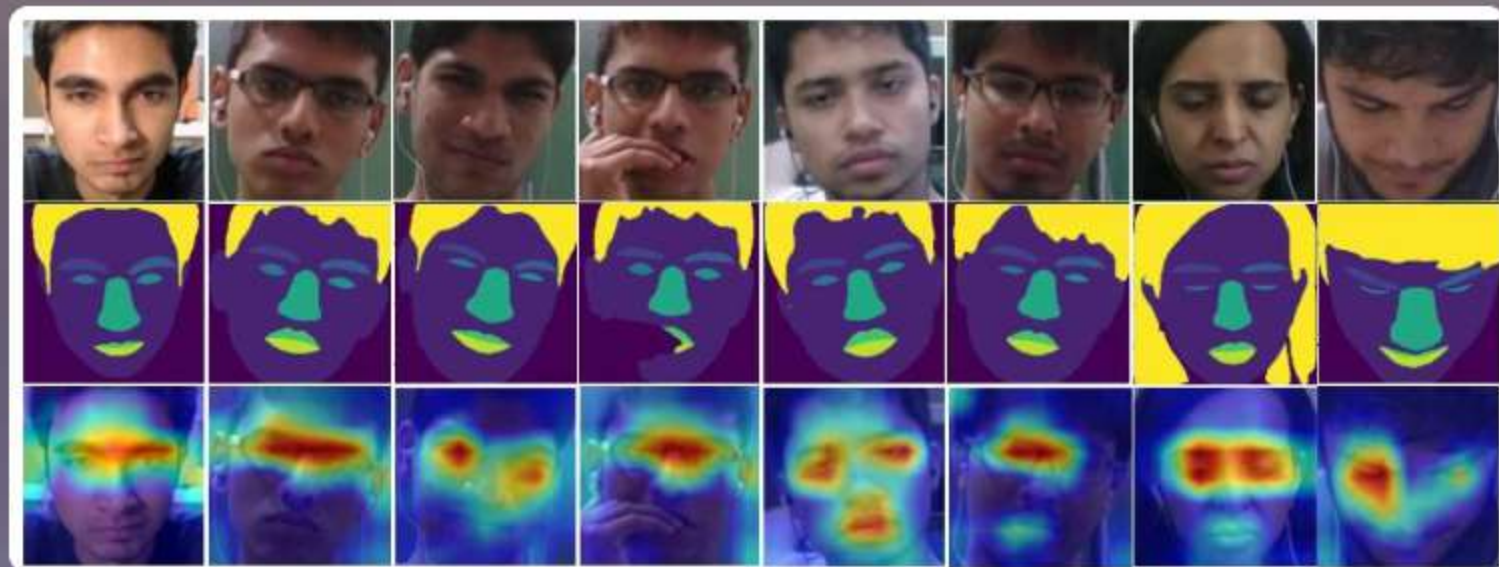
- Removing facial masking, reconstruction loss, or adversarial loss reduces accuracy.
- Best results achieved when all modules work together.

Interplay Effects of Masking and Loss Functions

Exhaustive experiments combining masking strategies with reconstruction and adversarial losses reveal:

- Combined losses outperform single loss functions by up to 2.15% accuracy improvement.
- Facial masking combined with both losses yields the best performance.
- Design of FMAE effectively captures key facial information and reconstructs masked regions efficiently.





Visualization Analysis of Facial Parsing and Feature Maps

Facial parsing effectively segments regions like glasses and occlusions, supporting mask operations. Grad-CAM visualizations show FMAE focuses on key facial areas such as eyes, even under occlusion and non-frontal poses, demonstrating robust facial representation learning.

Extended Experiments: Facial Expression Recognition

Datasets

Evaluated on AFEW, CREMA-D, and RAVDESS datasets covering basic expressions like anger, happiness, fear, and neutral.

Results

FMAE achieves competitive accuracy: 62.71% on AFEW (+3.29% improvement), 71.58% on CREMA-D, and 74.83% on RAVDESS without using audio data, demonstrating robust facial feature learning.



Conclusion and Future Directions

FMAE is an efficient self-supervised model that learns rich facial representations for student engagement recognition, outperforming supervised methods. It introduces facial masking and reconstruction modules to enhance learning.

Future work includes optimizing model size for lightweight deployment and addressing privacy and ethical concerns in online learning engagement detection.