A Self-Supervised Learning Network for Student Engagement Recognition From Facial Expressions

Wen-Long Zhang, Rui-Sheng Jia, Hu Wang, Cheng-Yue Che, and Hong-Mei Sun

***Abstract*— Student engagement in online learning is an important indicator for measuring learning effectiveness. Due to the fact that facial video data of students during online learning contains a wider range of information such as time, current research has begun to focus on obtaining student engagement from video data. These studies primarily rely on supervised learning methods and have achieved certain success. However, the longstanding lack of large-scale and high-quality labeled data, as well as the time-consuming and laborious sample labeling work, have to some extent hindered their further improvement. To solve this problem, this paper proposes a self-supervised learning method, Facial Masked Autoencoder (FMAE), which is used to construct a student engagement recognition model. This method uses a masked autoencoder to process a large number of unlabeled facial videos, and performs self-supervised pre-training by learning masked facial features from the reconstruction process. In order to promote the encoder to better mask learning for the face, a new facial mask strategy and reconstruction module have been proposed. With this method, the model can not only focus on important facial regions, but also obtain more accurate appearance features and spatio-temporal details. Experiments have demonstrated that the proposed method achieves excellent results on DAiSEE and EmotiW datasets, showing its potential in the task of student engagement recognition.**

***Index Terms*— Online learning, student engagement, self-supervised learning, masked autoencoder.**

I. Introduction

NOWADAYS, online education has attracted wide attention. With its flexibility and portability, it can provide rich educational resources, and allow students to experience national or even global quality teaching resources without going out [1]. However, there are some limitations to this form of education, one of which is the lack of feedback and interaction. In traditional classroom education, teachers usually adopt some corresponding teaching methods, such as observing students' facial expressions and other behaviors to grasp student engagement in real time. However, in online education, face-to-face communication between students and teachers is often lacking [2]. Therefore, the ability to assess student engagement and provide timely feedback or intervention in online education is an important factor that affects students' learning outcomes and ensures their learning effects.

The identification methods of student engagement can be divided into two categories according to the type of input data: the identification methods of student engagement based on static data and the identification methods of student engagement based on dynamic data. The former uses static images as input, while the latter is designed to identify student engagement in dynamic image sequences or videos. Since methods based on static data ignore the critical temporal information of the face, this paper will focus on methods based on dynamic data. This method has higher accuracy and can better reflect student engagement.

The identification of student engagement based on dynamic data is mainly relies on supervised learning methods. Currently, researchers have exploited diversified deep neural networks for this assignment, including 2D/3D convolutional neural networks (CNN) [15], [16], [17], ensemble neural networks [18], [19], [20], [21] and a more advanced Transformer based architecture [22], [23]. Despite the remarkable success of supervised learning methods in the task of student engagement identification, there are still several obstacles that limit the further development of this field. Firstly, the currently available datasets for student engagement recognition are rather limited, and the training samples within these datasets are relatively single (only contain videos of a few subjects and only cover a few scenes). Secondly, supervised learning methods are prone to overfitting and therefore have poor generalization ability when applied to other datasets or practical applications. Finally, the collection of large-scale and high-quality annotation data is a time-consuming and laborious task [3]. Considering that there are a large number of unlabeled facial videos on the Internet, the task of student engagement recognition can be accomplished by self-supervised learning methods (Fig. 1. illustrates the process of our self-supervised learning method in contrast with the supervised learning method).

Self-supervised learning has achieved remarkable achievement in multiple deep learning researches. Among them, MAE, as one of the important methods of generative self-supervised learning, has recently took out hitherto unknown results in numerous deep learning researches [4]. Initially, this method was applied to masked language modeling in the field

Manuscript received 6 May 2024; revised 4 July 2024 and 22 July 2024; accepted 26 July 2024. Date of publication 31 July 2024; date of current version 23 December 2024. This work was supported by the Humanities and Social Science Fund of the Ministry of Education of the People's Republic of China under Grant 22YJAZH036. This article was recommended by Associate Editor Z. Tang. *(Corresponding authors: Rui-Sheng Jia; Hong-Mei Sun.)*

The authors are with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China (e-mail: zhangwenlong@sdust.edu.cn; jrs@sdust.edu.cn; wanghu@sdust.edu.cn; chechengyue@sdust.edu.cn; skd991915@sdust.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3436029.

Digital Object Identifier 10.1109/TCSVT.2024.3436029

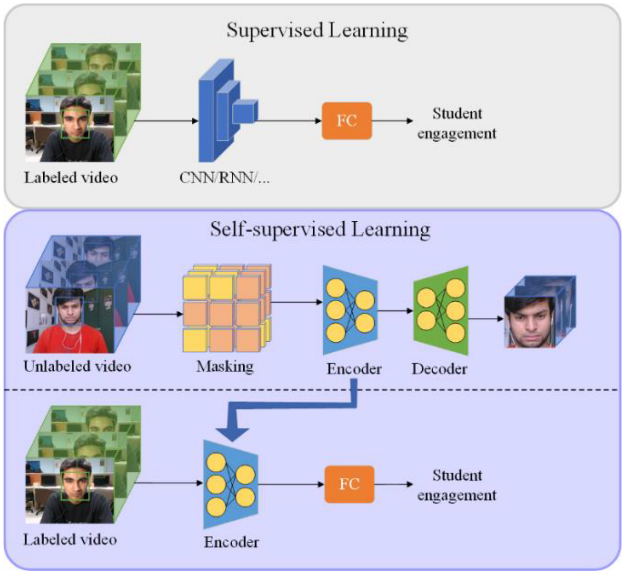

Fig. 1. Supervised and self-supervised learning methods.

of natural language processing, where models are pre-trained by a strategy of first masking and then reconstruction. With the success of BERT based mask methods [5], mask autoencoders have also been continuously explored in the visual field. In the realm of point cloud learning, McP-BERT has shown tremendous potential [6]. It has successfully optimized the process of self-supervised pre-training for point clouds by ingeniously introducing multi-choice token and utilizing the high-level semantics information learned from transformers. Additionally, VideoMAE extends MAE to the video fields and has achieved deep impression results on many universal video datasets [7]. In order to better perform the masking task, some studies have adopted a variety of different design options, for example pixel level masking [8], [9], token level masking [10], [11], depth feature based masking [12], and the use of ViT [13]. In addition, in order to better model the spatio-temporal patterns of input data, recent studies have also explored strategies such as masked motion modeling [14] and tube masking [7]. Inspired by these research strategies, this paper proposes a Facial Masked Autoencoder (FMAE) based on self-supervised learning, which learns masked facial features during the reconstruction process to solve the dilemma that supervised learning methods face in video-based student engagement recognition. The main contributions of this paper are summarized as follows:

1) A new self-supervised learning network is proposed for student engagement recognition. In this method, mask autoencoder is applied to process a large number of unlabeled facial video data, which effectively reduces the cost of sample annotation.
2) An encoder with a new masking strategy is designed, which forces the encoder to pay more attention to the spatio-temporal details of the specific facial regions through the mask operation of these regions in the time series, so as to improve the learning ability of facial features.
3) An effective reconstruction module is constructed, which helps the model accurately reconstruct the masked facial region through joint reconstruction loss and adversarial loss, thereby promoting the model to learn more detailed and rich facial representations during the reconstruction process.
4) The experimental results indicate that FMAE has achieved remarkable achievements on DAiSEE and EmotiW datasets, which proves its effectiveness in the task of student engagement recognition.

The structure of the remaining parts of this paper is as follows. The second section discusses the relevant research in the field of student engagement recognition. The third section describes the details of the proposed method in this paper. The fourth section presents the experimental results and analysis on DAiSEE and EmotiW datasets. The fifth section provides a summary of this paper and looks forward to future work directions and potential improvement space.

II. Related Works

In the research of using video data to obtain student engagement, researchers mainly adopt supervised learning methods and strive to develop more advanced deep learning architectures to extract valuable spatial-temporal information from original facial videos, so as to obtain student engagement. In general, three trends can be summed up.

Firstly, one trend is to straight utilize 3D CNNS to obtain combined spatio-temporal features from original face videos. Geng et al. [15] used C3D model in their study to identify student engagement by modeling appearance (facial expression) and motion information in videos. Zhang et al. [16] innovatively introduced I3D, an excellent network in the field of behavior recognition, into the field of student engagement recognition. The structure of the model is improved and optimized according to the characteristics of the student engagement recognition dataset, so that it could accurately analyze and evaluate student engagement from facial video data, thus significantly improving the accuracy of student engagement recognition. In addition, Mehta et al. [17] based on 3D DenseNet and self-attention mechanisms designed a neural network model to identify and assess student engagement and emotional state in online education. The self-attention block in the model helps to extract enhanced facial features (spatial features, temporal features, spatial-temporal features), which enables the model to effectively detect the student engagement status in the video sequence, ultimately achieving satisfactory accuracy results.

Secondly, the second trend is to use an ensemble model, first using a 2D CNN to obtain facial features from each static frame, and then using RNN to synthesize the dynamic temporal information of all frames. Wu et al. [18] designed a feature-based method that they obtain facial features and the upper part of the body features from videos through 2D convolutional neural networks, and combined Long Short-Term memory (LSTM) and Gated Recurrent Unit (GRU) to classify the extracted features to identify the degree of student engagement. In addition, Zhu et al. [19] designed a GRU

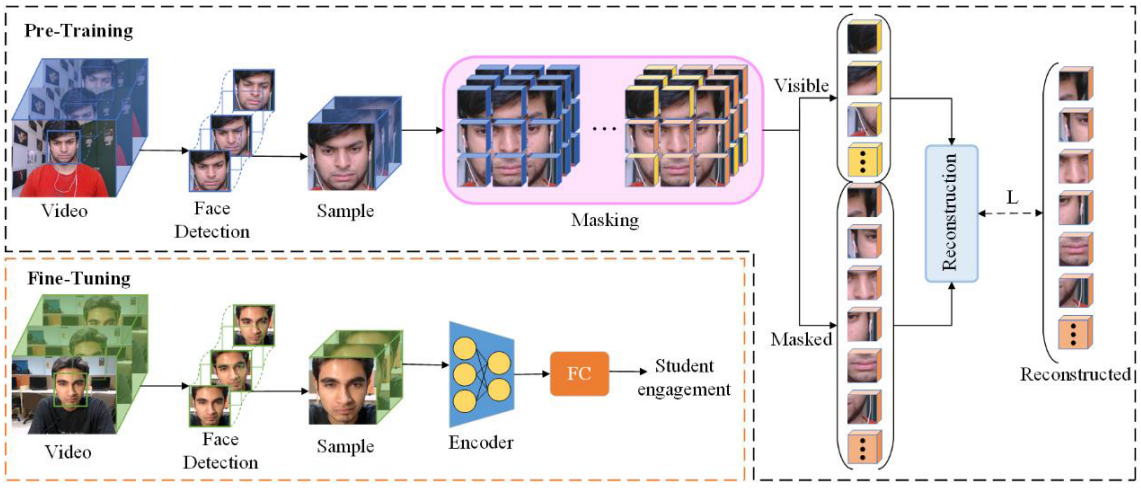

Fig. 2. Overview of the FMAE architecture. It consists of two phases: self-supervised pre-training and fine-tuning. Pre-training architecture: FMAE learns facial representation from unlabeled network video data in a self-supervised learning manner. Fine-tuning architecture: The pre-trained model is fine-tuned to make it suitable for the student engagement recognition task.

model for attention mechanism, which extracts facial features, posture features and CNN features from videos. The final student engagement level is predicted by combining these three different types of features. Additionally, Liao et al. [20] proposed a deep face spatio-temporal network (DFSTN) for online learning, which uses a pre-trained SE-ResNet-50 to extract the spatial features of the face and uses an LSTM with global attention to generate attention hidden states, so as to detect student engagement. In addition, Abedi and Khan [21] treated detecting student engagement as a classification mission related to spatio-temporal information and proposed a new end-to-end ResNet + TCN hybrid neural network. Among them, ResNet is used to extract spatial features of faces from successive video frames, while TCN obtains the degree of engagement by processing temporal changes of faces in video frames.

Recently, with the rise of Transformer, several studies have begun to exploit its global dependency modeling capabilities for better performance, which forms a third trend. Xusheng et al. [22] designed a new class attention method based on ViT [13], called CavT. In this method, the video sequence is divided into blocks along temporal and spatial dimensions, each block contains the same part of the adjacent image, and the blocks are converted into patch embeddings using linear projection. Finally, ViT with a class attention module is used to process these patch embeddings. Through unified training of long videos of variable length and short videos of fixed length, this method can carry out student engagement learning and achieve good accuracy. Chen et al. [67] proposed an innovative Multi-relation Perception Network (MRAN), which deeply explores and learns multi-level relationships among local regions, between global-local features, and among different samples by comprehensively focusing on the significant features of the whole face and local areas. With this multi-dimensional feature capture, MRAN is able to extract richer facial features from different angles. Additionally, Qin et al. [23] proposed a multi-task network known as SwinFace, which is based on the Swin Transformer and capable of performing a variety of tasks including face recognition, expression recognition, age estimation, and facial attribute estimation. To address potential conflicts between tasks and to meet their unique requirements, they have integrated a Multi-Level Channel Attention (MLCA) module into each subnet. This module can adaptively select the most appropriate feature levels and channels to accurately execute tasks. With this design, SwinFace not only demonstrates a deep understanding of facial features, ensuring exceptional accuracy in the recognition task of student engagement, but also excels across all tasks.

However, existing student engagement recognition methods mainly use supervised learning methods to analyze manually annotated data, so they are limited by existing student engagement datasets. Different from them, this paper designs a self-supervised learning method that could learn effective facial features from a quantity of unlabeled face video data and apply it to the field of student engagement recognition.

III. Proposed Method

A. FMAE

Considering the task of student engagement recognition as a whole, it can be analyzed from two different aspects. The first is the area associated with facial features, which includes various regions of the face (eyes, nose, mouth, etc.), mainly studying the facial shape and texture of these area; Secondly, temporal information in facial movements needs to be taken into account, so spatio-temporal modeling is very necessary. To achieve this goal, a new self-supervised learning framework FMAE is proposed, as shown in Fig. 2. The framework applies facial masking strategy to mask the unlabeled video data. Then, the masked image is reconstructed from the visible input using an encoder decoder architecture, and a reconstruction module is used to train the encoder between the reconstructed image and the input image, which is applied to the recognition task of student engagement.

FMAE mainly consists of two parts: facial mask strategy and reconstruction module. For a given training dataset $S =$

$\{V_i\}_{i=1}^{N}$, where N is the number of videos in the dataset and V is a single video in the input dataset. From the original input video V, the facial region is obtained by tracking and cropping, and then random time sampling is performed to obtain $v \in \mathbb{R}^{C\times T\times H\times W}$ (C, T, H, W respectively represent the number of channels of the processed video, the length of the video, and the height and width of a single frame of the video). Through the facial masking strategy, additionally map v into n visible tokens and $(p-n)$ masked tokens using a predefined masking ratio $r = (p-n)/p$. Visible tokens are represented as $K_v \in \mathbb{R}^{n\times d}$, masked tokens are represented as $K_m \in \mathbb{R}^{(p-n)\times d}$, where d is the embedded dimension, and p is the total number of tokens mapped from v, for a given token $p = T/t \times H/h \times W/w$ with $t \times h \times w$ dimension. Therefore, FMAE passes specific areas of the face through the above tokens for mask learning, which can better focus on important features of the face. The visible tokens are mapped through the encoder to the latent feature space z, which captures the key information of the facial representation. Then, the decoder will utilize the information in the latent feature space z to reconstruct it into $(p-n)$ masked tokens. The reconstruction process aims to recover the masked information in order to enhance understanding of facial features. It is quite a challenging task to reconstruct a spatio-temporal face from original pixels, so a reconstruction module is designed for better training.

The FMAE model learns rich and detailed face representations from face videos through self-supervised learning, and then fine-tuning [24] is used to make it suitable for the task of student engagement recognition. For a given dataset $S_e = \{v_j, y_j\}_{j=1}^{N}$ of student engagement, a linear fully connected layer with embedding parameters w is introduced to align the latent feature space obtained through the encoder with the label space of the video data of student engagement. In the fine-tuning, the backbone network is frozen and only the w parameters of the full connection layer are updated. This can effectively utilize the learned facial representations and associate them with labels of student engagement, thereby improving the performance of student engagement recognition.

B. Facial Masking Strategy

Since video can be regarded as the temporal extension of static appearance, and there is correspondence between video frames, such temporal correlation may lead to information leakage, where masked content in one frame may be visible in another frame [25]. Therefore, in order to make better use of spatio-temporal information of video data, a facial mask strategy is designed inspired by strategies such as tube mask [7]. The architecture of this strategy is shown in Fig. 3. Specifically, the strategy processes each spatio-temporal cube by dynamic tracking and masking, which means that the same facial region is masked at the spatial location of different time frames of the video, thus maintaining consistency while shielding the influence of correlation information. In this way, when performing later reconstruction tasks, the model needs to overcome the challenge introduced by the mask, which is to recover the complete facial representation from the visible

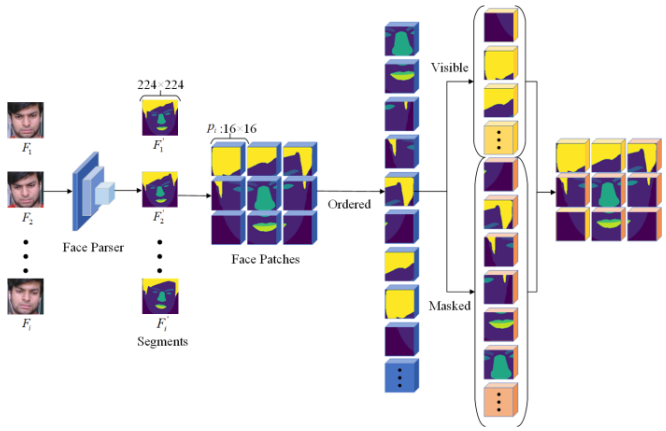

Fig. 3. Facial mask strategy.

part of the facial information. Therefore, through this masking strategy, the model is encouraged to learn the local and global features of the face, so as to better utilize the spatio-temporal information in the video data and improve the learning effect of the face representation.

Before the mask operation, face parsing is performed first [26]. Through face parsing, the facial image can be divided into different regions, such as left eye, right eye, nose, mouth, hair, skin and background, as shown in Equation (1). Based on the characteristics of these facial regions, mask strategies that are more suitable for faces can be designed.

$$s = \{f_{le}, f_{re}, f_{no}, f_{mo}, f_{ha}, f_{sk}, f_{bg}\} \tag{1}$$

Firstly, different priorities are set for different facial regions according to their importance and attention, as shown in Equations (2) and (3). In general, the left eye, right eye, nose, mouth and other regions of the face provide more abundant and important facial feature information, so these facial regions are set higher priority. Next, using a predefined masking ratio, the mask operation is first applied to the facial regions with higher priority. This means that the information of these regions will be hidden during the encoding and reconstruction process, thus ensuring the accurate acquisition and reconstruction of important facial features. Subsequently, mask operations are performed on areas such as hair, skin, and background. With this selective mask operation, it is possible to reduce the reconstruction of irrelevant or secondary facial information, thereby improving the quality and accuracy of facial reconstruction.

$$S_{high} = \{f_{le}, f_{re}, f_{no}, f_{mo}\} \tag{2}$$

$$S_{low} = \{f_{ha}, f_{sk}, f_{bg}\} \tag{3}$$

By adopting this masking strategy, map the input v into n visible tokens and $(p-n)$ masked tokens. Next, these visible tokens are used to encode and reconstruct the spatio-temporal variations of the face. Facial mask provides a more efficient masking strategy for facial reconstruction tasks, which enables more accurate acquisition of facial appearance features and spatio-temporal details. Algorithm 1 shows the processing of the facial mask operation.

Algorithm 1 Facial Masking Procedure

Require: $v \in \mathbb{R}^{C\times T\times H\times W}, r$
$regionMap \leftarrow FaceParser\ (v)$ ▷ $Face-Parsing$
$regionMap \in \{f_{le}, f_{re}, f_{no}, f_{mo}, f_{ha}, f_{sk}, f_{bg}\}$
$s = \{f_{le}, f_{re}, f_{no}, f_{mo}, f_{ha}, f_{sk}, f_{bg}\}$ ▷ $Prioritize\ Regions$
$p = T/t \times H/h \times W/w$ ▷ $tokens\ for\ each\ v$
▷ ($3D\ cube\ tokens\ have$ dim $ension\ of\ t \times h \times w\ each$)
$(p-n) \leftarrow r \times p$ ▷ $Number\ of\ masked\ tokens$
$K_m \leftarrow \{\}$ ▷ $Initializemaskedtokens$
$S_{high} = \{f_{le}, f_{re}, f_{no}, f_{mo}\}$
$S_{low} = \{f_{ha}, f_{sk}, f_{bg}\}$ ▷ $Set\ priority$
$for\ s'\ in\ s\ do$
  $K_m \leftarrow \{s'\}$
  $if\ len\ (K_m) == (p-n)\ then$
    $break$
  $end\ if$
$end\ for$
$K_v \leftarrow K - K_m$ ▷ $K\ is\ all\ tokens\ from\ v$

Algorithm 2 Training procedure for FMAE

Require: $Ft, En\ (Encoder), De\ (Decoder), D, S, r, w, S_e$
$while\ not\ converged\ do$ ▷ $FT-MAE\ pre-training$
  $v \leftarrow S$ ▷ $sample\ batch$
  $\{K_m, K_v\} \leftarrow Ft\ (v, r)$
  ▷ $Facial\ Trajectory\ Masking\ (See\ Algorithm1)$
  $K'_m \leftarrow DeEn\ (K_v)$
  $\{D\} \leftarrow \nabla_{\{D\}}\mathcal{L}^d\left(K_m, K'_m\right)$
  $K'_m \leftarrow DeEn\ (K_v)$
  $\{En, De\} \leftarrow \nabla_{\{En,De\}}\mathcal{L}^g\left(K_m, K'_m\right)$
$end\ while$
$while\ not\ converged\ do$
  $\{v, y\} \leftarrow S_e$ ▷ $sample\ batch$
  $K \leftarrow v$ ▷ $K\ is\ all\ tokens\ from\ v$
  $y' \leftarrow wEn\ (K_v)$
  $\{w\} \leftarrow \nabla_{\{w\}}\mathcal{L}_e\left(y, y'\right)$ ▷ $LinearProbing$
$end\ while$

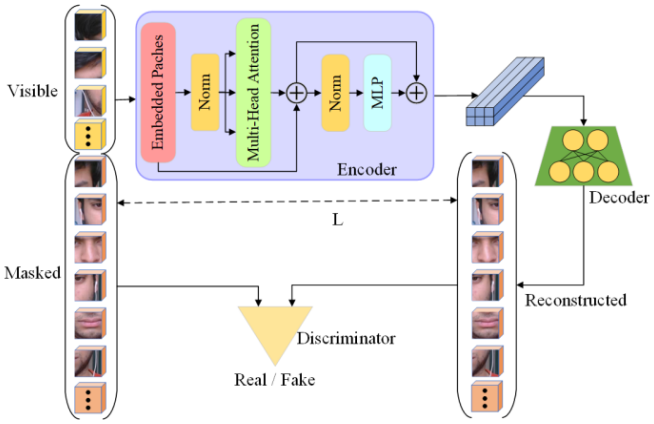

Fig. 4. Reconstruction module.

C. Reconstruction

After the facial masking, the facial reconstruction is realized through the cooperative work of the encoder and decoder. The architecture of the reconstruction module is shown in Fig. 4. The n visible tokens obtained after the mask operation are entered to the encoder, and the encoder maps these visible tokens to the latent feature space z. The visible tokens are used as a guideline to manufacture the masked region of the face, which enables the decoder to reconstruct the masked tokens from the latent feature space. In the reconstruction module, the encoder and decoder adopt the identical architecture as the ViT [13]. In order to guide the training of the reconstruction process, the reconstruction loss [27] and adversarial loss [28] are introduced to help the model learn how to accurately reconstruct the masked facial region by combining these two loss functions between the reconstructed cube and the masked cube. Algorithm 2 outlines the training process of the FMAE method. FMAE mainly facilitates model training by introducing reconstruction loss and adversarial loss.

Reconstruction loss. The reconstruction loss aims to minimize the difference between each pixel of the generated image and the original image, helping the model generate results that are more accurate and closer to the target. For the input visible tokens K_v, its visible facial information is utilized by the mask autoencoder in the reconstruction module to reconstruct the masked facial region K'_m. In order to better complete the process, the weight in the mask autoencoder is updated by minimizing the mean square error loss in spatio-temporal facial patterns. Reconstruction loss is defined as follows:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}\left\|K_m^{(i)} - K'^{(i)}_m\right\|_2 \tag{4}$$

where N is the total number of videos in the input dataset S, $K_m^{(i)}$ and $K'^{(i)}_m$ are the masked tokens and the reconstructed tokens of the i-th video data in S.

It should be noted that since the reconstruction loss is calculated based on the difference between image pixels, its use may lead to the reduction of the average pixel error of the whole image, which may cause image blurring and information loss. To better perform the reconstruction task, the model is trained by joint adversarial loss.

Adversarial loss. In order to better reconstruct the masked spatio-temporal facial region and learn richer and more effective feature representation, adversarial loss is introduced in the reconstruction module. Adversarial loss is based on the idea of generative adversarial network, by training a discriminator to evaluate the authenticity of the reconstructed image, while training the reconstruction module to deceive the discriminator to the maximum extent. This process of adversarial training can help the reconstruction module learn to generate more realistic and clearer reconstruction results, so as to compensate for the blurring and information loss problems that may be caused by the reconstruction loss. Adversarial loss is defined as follows:

$$\mathcal{L}_a^d = \frac{1}{N}\sum_{i=1}^{N}\left[\log D\left(K_m^{(i)}\right) + \log\left(1 - D\left(K'^{(i)}_m\right)\right)\right] \tag{5}$$

$$\mathcal{L}_a^g = \frac{1}{N}\sum_{i=1}^{N}\log\left(1 - D\left(K'^{(i)}_m\right)\right) \tag{6}$$

where $\log D\left(K_m^{(i)}\right)$ is the probability that the discriminator determines the real data (i.e., masked tokens) as real data, and $\log\left(1 - D\left(K'^{(i)}_m\right)\right)$ is the probability that the discriminator determines the false data (i.e., reconstructed tokens) as false

Fig. 5. Examples from the DAiSEE dataset: the first row represents high engagement, the second row represents engagement, the third row represents low engagement, and the fourth row represents disengagement.

Fig. 6. Examples from the EmotiW dataset: the first row represents high engagement, the second row represents engagement, the third row represents low engagement, and the fourth row represents disengagement.

data. The overall loss formula in this framework is as follows:

$$\mathcal{L}^g = \mathcal{L}_r + \mathcal{L}_a^g \tag{7}$$

$$\mathcal{L}^d = \mathcal{L}_a^d \tag{8}$$

Thus, better results can be achieved in the reconstruction task by using reconstruction loss and adversarial loss in combination. The reconstruction loss ensures the accuracy and consistency of the reconstructed image, while the adversarial loss provides an additional optimization mechanism to make the reconstructed results clearer and more realistic.

IV. Experiments

In this section, experiments are performed on the DAiSEE and EmotiW datasets to evaluate the proposed model. Before presenting the results, the experimental setup is first described, including the dataset, evaluation metrics, and experimental details. Afterwards, the proposed method is compared with the current SOTA method. Finally, a series of method analysis and ablation studies are carried out on the proposed method.

A. Experimental Settings

1) Dataset:

a) DAiSEE: This dataset consists of videos from 112 online learners, with a total of 9068 video samples. [29]. The videos were labeled according to the four states of the learners when watching the online course, including boredom, confusion, frustration, and engagement. Each state is divided into four levels: level 0(very low), level 1(low), level 2(high), and level 3(very high). The focus of this paper is to classify the degree of student engagement in online learning. The length of each video is 10 seconds, the frame rate is 30 frames per second, and the resolution is 640 × 480 pixels. Fig. 5. shows sample examples of different categories in the dataset. In our experiments, these datasets are used to fairly compare the previous methods with the proposed method. The final results report the performance on 1784 test videos.

b) EmotiW: This dataset is provided for measuring student engagement in the sub challenge of EmotiW [30]. The dataset contains videos of 78 people (25 females and 53 males, aged from 19 to 27 years) during online learning. There are 262 videos in total, which include 148 training videos, 48 validation videos and 67 test videos. The videos have a resolution of 640 × 480 pixels, 30 frames per second, and the length of each video is approximately 5 minutes. The level of engagement for every video is divided into four values corresponding to the lowest to highest engagement levels, where 0 indicates disengagement at all, 0.33 indicates low engagement, 0.66 indicates engagement, and 1 indicates high engagement. Fig. 6. presents some examples of samples from different categories in the dataset. In this subchallenge, only the training set and the validation set are publicly available, and we use the validation set to validate the proposed method.

2) Evaluation Metrics:

a) Accuracy: In this work, accuracy [31] is expressed as the number of correctly classified samples divided by the sum of the number of positive (correctly classified) and negative (misclassified) samples of all test samples, expressed by the formula as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively.

b) MSE: Mean Squared Error (MSE) is a commonly used measure of the difference between the predicted value of the model and the actual observed value, which measures the squared average distance between the true value of the data and the predicted value of the model. MSE is defined as follows:

$$MSE = \frac{1}{B}\sum_{i=1}^{B}\left(y_i - \widehat{y_i'}\right)^2 \tag{10}$$

where B is the number of samples contained in a batch, y_i is the true value of the i-th sample, and y_i' is the predicted value of the i-th sample.

c) Precision: Precision measures the proportion of samples predicted to be positive in a task that are actually positive. The precision is defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

where TP and FP are respectively predicted as positive samples and actually are positive samples, while predicted as positive samples are actually negative samples.

d) Recall: The recall rate measures the proportion of samples predicted to be positive out of the actual positive samples in the task. Recall is defined as follows:

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

e) F1-Score: F1-Score take into account both precision and recall, and is used to measure the overall performance of the task. F1-Score are defined as follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{13}$$

f) Implementation details: Firstly, for a given facial video, there is a high degree of redundancy between consecutive frames. Therefore, in order to consider meaningful frames in the temporal direction, a minimum temporal stride of 2 is used for frame sampling. Given an input video of size $3\times16\times224\times224$, generate $8\times14\times14$ spatio-temporal patterns of size $2 \times 16 \times 16$. When applying a facial mask strategy, FMAE performs mask operations on these spatio-temporal cubes using a predefined masking ratio. Relevant studies have shown that 90% mask is more suitable for model training [7], [42]. The goal of FMAE is to produce masked tokens from fewer visible tokens. After the masking operation, each token is mapped via the encoder to a latent feature space of dimension 768. Based on this latent feature space, the masked facial spatio-temporal cube is reconstructed. ViT-B is used as the encoder backbone. Some of the hyperparameters in the pre-training are as follows: $lr = base\ learningrate \times batch\ size/256$, relative to the overall batch size, the basic learning rate is linearly scaled [32]. AdamW optimizer [33], basic learning rate $1.5e-4$, momentum $\beta_1 = 0.9, \beta_2 = 0.95$ and learning rate scheduler with cosine decay are used for self-supervised pre-training [34]. In regard to fine-tuning process, the linear probing method is used, using Adam optimizer [35], $\beta_1 = 0.5$, $\beta_2 = 0.9$, the base learning rate is $1e-4$ and the weight decays to 0.

B. Performance Comparison

Firstly, FMAE is compared with the most advanced supervised learning methods on DAiSEE and EmotiW datasets in Table I and Table II respectively. On the DAiSEE dataset, FMAE outperforms the previous best methods (i.e., ResNet + TCN and Optimized ShuffleNet v2) with significant accuracy, achieving a noteworthy improvement of 0.84%. As for similar observations on the EmotiW dataset, in this experiment, the last fully connected layer is reshaped to adapt the model to the regression task. The FMAE method also achieves a significant MSE, outperforming most supervised learning methods and achieving comparable performance (0.0112 difference) to the current best method (i.e., MAGRU). The experiments in Table I and Table II show that FMAE can learn powerful facial representation and is well applied to the task of student engagement recognition.

TABLE I
PERFORMANCE COMPARISION ON DAISEE DATASET

Dataset	Model	Accuracy (%)
DAiSEE	C3D [15]	48.10
	I3D [16]	52.40
	LRCN [36]	57.90
	DFSTN [20]	58.84
	C3D + TCN [21]	59.97
	DERN [37]	60.00
	ResNet + LSTM [21]	61.50
	3D DenseAttNet [17]	63.59
	ResNet + TCN [21]	63.90
	Optimized ShuffleNet v2[38]	63.90
	ours	64.74

TABLE II
PERFORMANCE COMPARISION ON EMOTIW DATASET

Dataset	Model	MSE
EmotiW	Dhall et al. (Baseline) [39]	0.1
	ResNet + TCN [21]	0.096
	C3D [15]	0.0904
	DenseAttNet [17]	0.0877
	Swin-L [40]	0.0813
	I3D [16]	0.0741
	DFSTN [20]	0.0736
	CavT [22]	0.0667
	MAGRU [19]	0.0517
	ours	0.0629

In addition, a comparison of the confusion matrix of FMAE on the DAiSEE dataset and state-of-the-art supervised learning methods is given in Fig. 7. It is worth noting that due to the highly unbalanced distribution of samples in the DAiSEE dataset, the method adopted in Fig. 7. (a)-(d) fails to perform a good classification of disengagement and low engagement samples. Fig. 7. (e)ResNet + TCN method, due to the adoption of customized sampling strategy and weighted loss, some samples of disengagement and low engagement can be correctly classified, but the classification of engagement and high engagement is affected, which makes the overall recognition effect worse. Fig. 7. (f) FMAE method, compared with other methods, the number of samples in the main diagonal of the confusion matrix is significantly increased. This result indicates that FMAE has learned rich and detailed facial representation in the pre-training process, which makes the recognition effect of different student engagement significantly improved.

C. Ablation Study

Extensive ablation studies are conducted on the DAiSEE and EmotiW datasets to show the effectiveness of each component.

1) Masking Ratio: In order to thoroughly investigate the specific impact of different masking ratios on model performance, a series of experiments are carried out on two datasets DAiSEE and EmotiW, which select multiple different values of

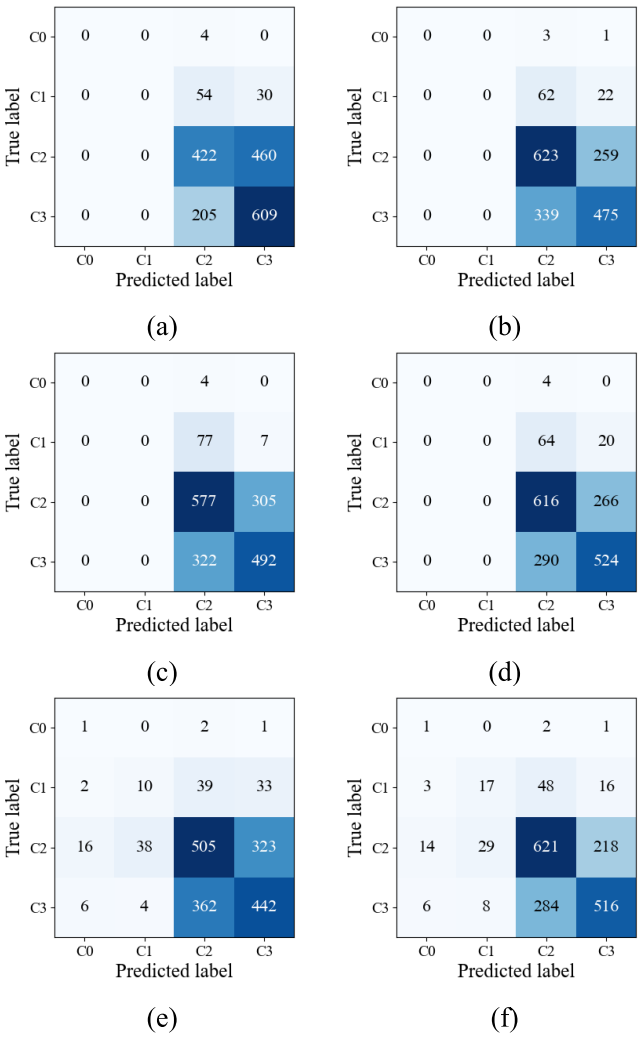

Fig. 7. Confusion matrices of different methods applied on the DAiSEE dataset (a) C3D fine tuning [15], (b) ResNet + LSTM [21], (c) C3D + TCN [21], (d) ResNet + TCN [21], (e) ResNet + TCN with weighted sampling and weighted loss [21], (f)FMAE (ours).

masking ratio in the range of [0.25,0.95]. The purpose of this experiment is to provide a clear understanding of the effect of masking ratio on experimental performance to enhance feature understanding while maintaining the efficiency of the reconstruction task. As can be seen from Fig. 8, a masking ratio of 90% is the optimal masking ratio for FMAE. When the masking ratio is smaller, the reconstruction task is able to obtain more information, which reduces the ability of feature understanding. If the masking ratio is set too large, especially above 90%, the reconstruction task becomes extremely challenging, resulting in a lack of sufficient information to complete accurate reconstruction and learn detailed features, which makes the overall performance degrade. Therefore, after experimental verification, 90% is consistently chosen as the optimal masking ratio in all experiments to ensure that sufficient information could be obtained while maintaining high feature comprehension in the reconstruction task.

2) Minimum Temporal Stride: The exploration is conducted on the DAiSEE dataset to evaluate the performance of student

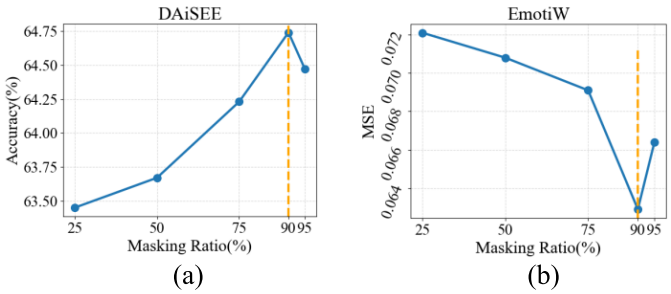

Fig. 8. Impact of Masking Ratio. Comparing the impact of different masking ratios on engagement recognition in the DAiSEE and EmotiW datasets.

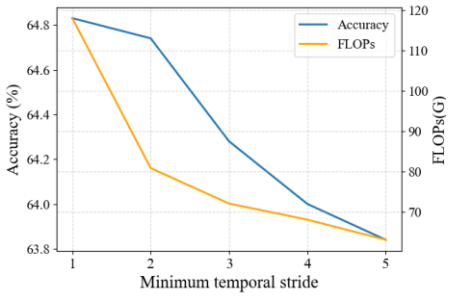

Fig. 9. Effect of the minimum time sampling stride.

engagement recognition tasks under five different settings of minimum temporal sampling strides. This experiment delved into the delicate balance between recognition accuracy and computational complexity, aiming to determine the optimal minimum temporal sampling stride. According to the trade-off relationship between recognition accuracy and computational complexity under various stride size settings in Fig. 9, it can be found that when the minimum sampling stride size is set to 2, the accuracy of student engagement recognition only shows a slight decrease compared with the more refined stride size, but the computational complexity is significantly reduced at this time. This indicates that the minimum sampling stride size of 2 is the most reasonable choice, which not only ensures the accuracy of identification, but also has a more efficient performance.

3) Masking Strategies: This ablation experiment aims to compare the differences between the proposed facial mask strategy and the existing mask strategies, namely frame mask, random mask and tube mask strategies. The frame masking strategy [41] is to partially mask the frame data, and the random masking strategy [42] is to randomly select some image regions for masking. The experiments in Table III show that these two methods are relatively less effective. This is due to the inherent relationship between adjacent frames, which is not processed by the two masking methods, resulting in information leakage in the mask and reconstruction process, which ultimately affects the accuracy of experimental results. Although the tube masking strategy [7] masks the same region of different frames, the mask region of this strategy is randomly selected. In contrast to the proposed face masking strategy, it does not select the regions in the face that contain important information for masking. Therefore, the experimental results of the tube mask are also slightly inferior.

TABLE III

ABLATION STUDIES OF DIFFERENT MODULES. FS: FACIAL STRATEGY, RL: RECONSTRUCTION LOSS, AL: ADVERSARIAL LOSS

Masking Strategy	DAiSEE				EmotiW
	Accuracy	Precision	Recall	F1-Score	MSE
Random	59.17	0.57	0.58	0.58	0.0748
Frame	57.51	0.56	0.57	0.56	0.0783
Tube	62.98	0.60	0.62	0.61	0.0671
Facial	64.74	0.63	0.65	0.64	0.0629

TABLE IV

ABLATION STUDIES OF MASKING STRATEGIES

Modules	DAiSEE				EmotiW
	Accuracy	Precision	Recall	F1-Score	MSE
FMAE	64.74	0.63	0.65	0.64	0.0629
w/o FS	62.98	0.60	0.62	0.61	0.0671
w/o RL	63.64	0.60	0.63	0.61	0.0668
w/o AL	63.79	0.61	0.64	0.62	0.0664
w/o FS&RL	60.83	0.56	0.59	0.58	0.0769
w/o FS&AL	61.69	0.58	0.61	0.59	0.0758

4) Different Modules: Keep other components fixed and remove some modules from the framework to verify the validity of the corresponding modules. Firstly, the validity of the facial masking strategy is verified. From Table IV, it can be observed that the accuracy of the model decreases significantly when the facial masking strategy is not used. This is mainly attributed to the fact that this strategy masks the regions containing important information in the face, forcing the model to learn these important facial features, so it makes a significant contribution to improving the accuracy of the model. Secondly, the reconstruction loss and adversarial loss in the model are removed respectively to verify the effectiveness of their separate effects. According to the experimental results in Table IV, when the reconstruction loss is not used, the accuracy of the model decreases by 1.1%. This may be because in the absence of fine-grained control of reconstruction loss, only the use of adversarial loss cannot guide the reconstruction process of the model at a more detailed level, thus affecting the performance of the model. Similarly, the accuracy of the model is also reduced when adversarial loss is not used. This indicates that adversarial loss plays an important role in model training, which enables the model to learn richer and more detailed facial representations through adversarial optimization mechanism. When the reconstruction loss and adversarial loss are combined in FMAE and applied to the training process of the model, the model shows the best performance. This indicates that the combination strategy of the two loss functions can better guide the model to accurately reconstruct the masked facial region, so as to improve the performance of the model. Finally, the facial masking strategy and different loss modules in the model are removed at the same time, so as to verify the effectiveness of the two modules acting simultaneously. According to the accuracy results shown in Table IV, the accuracy is greatly reduced and it can be concluded that the best results can only be obtained when all modules work together.

5) Interplay Effects: In this part, more in-depth ablation experiments are carried out to further verify the interaction between the masking strategy and the combination of different modules. In this ablation study, four masking strategies are combined with different loss functions in the reconstruction module through exhaustive experiments. From the experimental data in Table V, the following observations can be obtained: 1) When the same masking strategy is used, the experimental data show that the combined reconstruction loss and adversarial loss methods are generally superior in performance to those relying on a single loss function. Specifically, the performance improvement achieved by this combination strategy is up to 2.15% in accuracy, which further confirms that the combined design of loss functions in the reconstruction module has a key effect on the performance. 2) When using the same loss function, the combined approach with the facial masking strategy showed a significant performance improvement. This achievement is mainly attributed to the unique design of the facial masking strategy, which enables the model to focus more on the information of key facial regions. This result is consistent with the ablation results of masking strategy, which further confirms the important role of facial masking strategy in the optimization of model performance. 3) The combination scheme using the facial masking strategy and the designed reconstruction module (including reconstruction loss and adversarial loss) achieves the best results in the experiment. This result demonstrates that the design of FMAE is superior in the acquisition of facial key information and efficient reconstruction, and is the best combination scheme.

TABLE V

ABLATION STUDIES OF THE INTERPLAY EFFECTS BETWEEN MASKING STRATEGIES AND VARIOUS MODULES. ACC.: ACCURACY, PRE.: PRECISION, REC.: RECALL, F1: F1-SCORE

Masking Strategy	Modules		DAiSEE				EmotiW
	RL	AL	Acc.	Pre.	Rec.	F1	MSE
Random	✓		58.26	0.56	0.58	0.57	0.0792
		✓	57.72	0.56	0.56	0.56	0.0802
	✓	✓	59.17	0.57	0.58	0.58	0.0748
Frame	✓		57.42	0.54	0.56	0.55	0.0825
		✓	57.36	0.53	0.56	0.55	0.0833
	✓	✓	57.51	0.56	0.57	0.56	0.0783
Tube	✓		61.69	0.58	0.61	0.59	0.0758
		✓	60.83	0.56	0.59	0.58	0.0769
	✓	✓	62.98	0.60	0.62	0.61	0.0671
Facial	✓		63.79	0.61	0.64	0.62	0.0664
		✓	63.64	0.60	0.63	0.61	0.0668
	✓	✓	64.74	0.63	0.65	0.64	0.0629

D. Visualization Analysis

In order to further verify the effectiveness of FMAE method, the facial parsing process and the learned facial feature map are visualized. As shown in Fig. 10., the level of engagement decreases from left to right. From the visualization results of facial parsing in the second line, it can be noticed that the method we adopted can effectively divide the facial region and distinguish areas such as glasses and hand occlusion, which provides strong support for the subsequent mask operation. In the third line of Fig. 10., the results of a visualization experiment with Grad-CAM [43] are shown, from which it

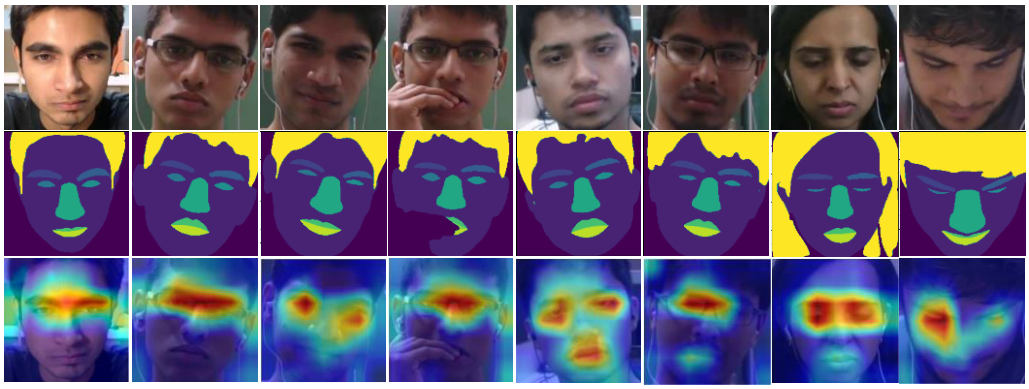

Fig. 10. Facial parsing and Grad - CAM visualization. The facial frame images obtained by cropping and other processing in the original video in the first row. The second row of face parsing results is visualized. The third row is visualized by Grad-CAM.

can be observed that the FMAE method is able to learn meaningful face representations and pay attention to key face areas. In addition, since the non-frontal poses also contain rich facial information related to student engagement, it can be noted in the third row that even in the presence of occlusion and non-frontal poses, the method is still able to focus on key areas such as eyes, thus obtaining a good facial representation.

E. Extended Experiments

FMAE learns rich and detailed facial representation from facial videos through self-supervised learning. In order to verify its ability to understand facial features, it is applied to facial expression recognition (FER) [44] task for further verification. Facial expression recognition refers to the classification of facial expressions by analyzing people's facial features. In order to evaluate the performance of FMAE in facial expression recognition task, experiments were conducted on AFEW [39], CREMA-D [45], and RAVDESS [46] datasets. The AFEW dataset is a collection of audio-video short clips gathered from movies and television series, consisting of 773 training samples and 383 validation samples. The dataset contains seven basic expressions (anger, disgust, fear, happiness, sadness, surprise, and neutral), and each video clip is labeled with a single expression category. CREMA-D is a high-quality audio-visual dataset consisting of 7442 video clips from 91 actors. The dataset covers six expressions, which are happy, sad, angry, fear, disgust, and neutral. RAVDESS is an audio-visual dataset of emotional speech and song. It consists of 2880 video clips from 24 professional actors with eight expressions (i.e., 7 basic expressions and calmness).

In the Table VI, the performance of FMAE in facial expression recognition task is compared in terms of accuracy. The results show that FMAE exhibits considerable competitiveness compared to SOTA methods. Specifically, on the AFEW dataset, FMAE achieved an accuracy of 62.71%, representing a performance improvement of 3.29% compared to the previous optimal model. This result indicates that FMAE is also competent for facial expression recognition tasks. Similarly, FMAE also shows considerable accuracy on both the CREMA-D and RAVDESS datasets. Although it is slightly weaker than the method of MSAF [60], CFN-SR [61] and Sinha et al. [54], this may be due to the fact that these methods leverage multimodal information from the data and provide more comprehensive and abundant decision information for facial expression recognition through other modalities, which makes these methods achieve more superior performance. Compared with these methods, FMAE still demonstrates good performance without using audio information. Through these results, it can be shown that FMAE learns robust, comprehensive and accurate facial features through the self-supervised way of mask autoencoder, which makes it competent for other face-related tasks.

TABLE VI
EXTENDED RESEARCH ON FACIAL EXPRESSION RECOGNITION TASKS

Dataset	Model	Accuracy (%)
AFEW	LBP-TOP (baseline) [47]	38.90
	Meng et al. [48]	51.18
	Kumar et al. [49]	55.17
	Li et al. [50]	54.30
	VGG13+VGG16+ResNet [51]	59.42
	ours	62.71
CREMA-D	Vougioukas et al. [52]	55.26
	Eskimez et al. [53]	65.67
	Sinha et al. [54]	75.02
	GRU [55]	55.01
	GAN [56]	58.71
	ViT [57]	67.81
	SepTr [58]	70.47
	ours	71.58
RAVDESS	AV-LSTM [59]	65.80
	MCBP [60]	71.32
	MSAF [60]	74.86
	CFN-SR [61]	75.76
	MMTM [62]	73.12
	ERANNs [63]	74.80
	ours	74.83

V. CONCLUSION

FMAE is an efficient self-supervised learning model that uses a large number of unlabeled face videos to cope with the dilemma of supervised learning methods in the task of student engagement recognition and promote its development. The model introduces two key designs, the facial masking strategy and the reconstruction module, to make the video reconstruction task more challenging, thus encouraging the model to learn more representative features. However, because the model adopts ViT architecture as the encoder, the number

of parameters of the model is large, which is not conducive to deployment in lightweight environment. The future research direction can focus on optimizing the parameter number of the model in order to provide the inference speed of the model and save computing resources. This can be achieved by compressing the model [64], model pruning [65], or designing more efficient architectures. In the process of promoting the development of student engagement recognition, privacy and ethical issues should also be taken seriously [66]. Future research should focus on how to identify student engagement in online learning while protecting student privacy.

REFERENCES

[1] Y. Cui et al., "A survey on big data-enabled innovative online education systems during the COVID-19 pandemic," *J. Innov. Knowl.*, vol. 8, no. 1, Jan. 2023, Art. no. 100295.

[2] S. K. Banihashem, O. Noroozi, P. den Brok, H. J. A. Biemans, and N. T. Kerman, "Modeling teachers' and students' attitudes, emotions, and perceptions in blended education: Towards post-pandemic education," *Int. J. Manage. Educ.*, vol. 21, no. 2, Jul. 2023, Art. no. 100803.

[3] Y. Wang et al., "FERV39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20890–20899.

[4] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2022, pp. 16000–16009.

[5] L. Kenton and L. K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 1, 2019, p. 2.

[6] K. Fu, M. Yuan, S. Liu, and M. Wang, "Boosting point-BERT by multi-choice tokens," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 438–447, Aug. 2023.

[7] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10078–10093.

[8] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[9] L. Zhang, X. Zhang, Q. Wang, W. Wu, X. Chang, and J. Liu, "RPMG-FSS: Robust prior mask guided few-shot semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6609–6621, Nov. 2023.

[10] S. Goley, R. Pradhan, and A. Welch, "Towards masked autoencoding pre-training for wide area motion imagery," in *Proc. Geospatial Informat. XIII*, Jun. 2023, pp. 51–64.

[11] P. Gao et al., "Mimic before reconstruct: Enhancing masked autoencoders with feature mimicking," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1546–1556, May 2024.

[12] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14648–14658.

[13] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[14] X. Sun et al., "Masked motion encoding for self-supervised video representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2235–2245.

[15] L. Geng, M. Xu, Z. Wei, and X. Zhou, "Learning deep spatiotemporal feature for engagement recognition of online courses," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 442–447.

[16] H. Zhang, X. Xiao, T. Huang, S. Liu, Y. Xia, and J. Li, "An novel end-to-end network for automatic student engagement recognition," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 342–345.

[17] N. K. Mehta, S. S. Prasad, S. Saurav, R. Saini, and S. Singh, "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement," *Appl. Intell.*, vol. 52, no. 12, pp. 13803–13823, 2022.

[18] J. Wu, B. Yang, Y. Wang, and G. Hattori, "Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 777–783.

[19] B. Zhu, X. Lan, X. Guo, K. E. Barner, and C. Boncelet, "Multi-rate attention based GRU model for engagement prediction," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 841–848.

[20] J. Liao, Y. Liang, and J. Pan, "Deep facial spatiotemporal network for engagement prediction in online learning," *Appl. Intell.*, vol. 51, pp. 6609–6621, Oct. 2021.

[21] A. Abedi and S. S. Khan, "Improving state-of-the-art in detecting student engagement with resnet and TCN hybrid network," in *Proc. 18th Conf. Robots Vis. (CRV)*, May 2021, pp. 151–157.

[22] X. Ai, V. S. Sheng, C. Li, and Z. Cui, "Class-attention video transformer for engagement intensity prediction," 2022, *arXiv:2208.07216*.

[23] L. Qin et al., "SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2223–2234, Apr. 2024.

[24] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15619–15629.

[25] H. Zhu, Y. Chen, G. Hu, and S. Yu, "Information-density masking strategy for masked image modeling," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2023, pp. 1619–1624.

[26] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K. K. Wong, "Progressive semantic-aware style transformation for blind face restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11891–11900.

[27] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by in painting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Oct. 2544, pp. 2536–2544.

[28] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–12.

[29] A. Gupta, A. D'Cunha, K. Awasthi, and V. Balasubramanian, "DAiSEE: Towards user engagement recognition in the wild," 2016, *arXiv:1609.0188*.

[30] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–8.

[31] S. Batra et al., "DMCNet: Diversified model combination network for understanding engagement from video screengrabs," *Syst. Soft Comput.*, vol. 4, Dec. 2022, Art. no. 200039.

[32] P. Goyal et al., "Accurate, large minibatch SGD: Training ImageNet in 1 hour," 2017, *arXiv:1706.02677*.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[34] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[36] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[37] T. Huang, Y. Mei, H. Zhang, S. Liu, and H. Yang, "Fine-grained engagement recognition in online learning environment," in *Proc. IEEE 9th Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2019, pp. 338–341.

[38] Y. Hu, Z. Jiang, and K. Zhu, "An optimized CNN model for engagement recognition in an E-Learning environment," *Appl. Sci.*, vol. 12, no. 16, p. 8007, Aug. 2022.

[39] A. Dhall, "EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 546–550.

[40] Z. Liu et al., "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.

[41] Y. Lee, H. Seong, and E. Kim, "Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1245–1253.

[42] C. Feichtenhofer, Y. Li, and K. He, "Masked autoencoders as spatiotemporal learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 35946–35958.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[44] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022.

[45] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.
[46] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north American English," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
[47] T. S. Ashwin, and R. M. R. Guddet, "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," *Future Gener. Comput. Syst.*, vol. 108, pp. 334–348, Jul. 2020.
[48] D. Meng, X. Peng, K. Wang, and Y. Qiao, "Frame attention networks for facial expression recognition in videos," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3866–3870.
[49] V. Kumar, S. Rao, and L. Yu, "Noisy student training using body language dataset improves facial expression recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 756–773.
[50] S. Li et al., "Bi-modality fusion for emotion recognition in the wild," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 589–594.
[51] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, Oct. 2016, pp. 433–436.
[52] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413, May 2020.
[53] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Trans. Multimedia*, vol. 24, pp. 3480–3490, 2022.
[54] S. Sinha, S. Biswas, R. Yadav, and B. Bhowmick, "Emotion-controllable generalized talking face generation," 2022, *arXiv:2205.01155*.
[55] A. Shukla, K. Vougioukas, P. Ma, S. Petridis, and M. Pantic, "Visually guided self supervised learning of speech representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6299–6303.
[56] G. He, X. Liu, F. Fan, and J. You, "Image2Audio: Facilitating semi-supervised audio emotion recognition with facial expression image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 3978–3983.
[57] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," 2021, *arXiv:2104.01778*.
[58] N.-C. Ristea, R. T. Ionescu, and F. S. Khan, "SepTr: Separable transformer for audio spectrogram processing," 2022, *arXiv:2203.09581*.
[59] E. Ghaleb, M. Popa, and S. Asteriadis, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2019, pp. 552–558.
[60] L. Su, C. Hu, G. Li, and D. Cao, "MSAF: Multimodal split attention fusion," 2020, *arXiv:2012.07175*.
[61] Z. Fu et al., "A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition," 2021, *arXiv:2111.02172*.
[62] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13286–13296.
[63] S. Verbitskiy, V. Berikov, and V. Vyshegorodtsev, "ERANNs: Efficient residual audio neural networks for audio pattern recognition," *Pattern Recognit. Lett.*, vol. 161, pp. 38–44, Sep. 2022.
[64] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," *Artif. Intell. Rev.*, vol. 53, no. 7, pp. 5113–5155, Oct. 2020.
[65] A. Onan, S. Korukoğlu, and H. Bulut, "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification," *Inf. Process. Manage.*, vol. 53, no. 4, pp. 814–833, Jul. 2017.
[66] G. Xu, H. Li, Y. Zhang, S. Xu, J. Ning, and R. H. Deng, "Privacy-preserving federated deep learning with irregular users," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 1364–1381, Mar. 2022.
[67] D. Chen, G. Wen, H. Li, R. Chen, and C. Li, "Multi-relations aware network for in-the-wild facial expression recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3848–3859, Sep. 2023.

Wen-Long Zhang was born in Shandong, China, in 1999. He received the B.S. degree from Qingdao University of Science and Technology, China, in 2022. He is currently pursuing the M.S. degree with Shandong University of Science and Technology. His research interests include image processing and deep learning.

Rui-Sheng Jia is currently a Full Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China, and the Leader of the Natural Science Foundation of Shandong Province, China. He has more than 30 first-author publications and has more than 50 co-author publications. His research interests include artificial intelligence, computer vision, information fusion, microseismic monitoring, and inversion.

Hu Wang was born in Anhui, China, in 1999. He received the B.S. degree from Qingdao University of Technology, China, in 2022. He is currently pursuing the M.S. degree with Shandong University of Science and Technology. His research interests include image processing and deep learning.

Cheng-Yue Che was born in Shandong, China, in 2001. She received the B.S. degree from Shandong University of Science and Technology, China, in 2023, where she is currently pursuing the M.S. degree. Her research interests include image processing and deep learning.

Hong-Mei Sun received the B.S. and M.S. degrees in computer science from Shandong University of Science and Technology, China, in 1995 and 2005, respectively. She is currently an Associate Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, and the Leader of the Key Research and Development Projects of Shandong Province, China. She has more than 20 first-author publications and has more than 50 co-author publications. Her research interests include computer vision, deep learning, and software engineering.