

Hierarchical Temporal Multi-Instance Learning for Video-based Student Learning Engagement Assessment

Jiayao Ma^{1,2,3*}, Xinbo Jiang^{1,2,3*}, Songhua Xu^{4†}, Xueying Qin^{1,2,3†}

¹School of Software, Shandong University, Jinan, China

²Engineering Research Center of Digital Media Technology, Ministry of Education, China

³Key Laboratory of Shandong Province for Software Engineering, China

⁴College of Engineering and Computing, University of South Carolina, Columbia, SC, USA
 mageayo123@gmail.com, xinboj@gmail.com, xus1@cec.sc.edu, qxy@sdu.edu.cn

Abstract

Video-based automatic assessment of a student’s learning engagement on the fly can provide immense values for delivering personalized instructional services, a vehicle particularly important for massive online education. To train such an assessor, a major challenge lies in the collection of sufficient labels at the appropriate temporal granularity since a learner’s engagement status may continuously change throughout a study session. Supplying labels at either frame or clip level incurs a high annotation cost. To overcome such a challenge, this paper proposes a novel hierarchical multiple instance learning (MIL) solution, which only requires labels anchored on full-length videos to learn to assess student engagement at an arbitrary temporal granularity and for an arbitrary duration in a study session. The hierarchical model mainly comprises a bottom module and a top module, respectively dedicated to learning the latent relationship between a clip and its constituent frames and that between a video and its constituent clips, with the constraints on the training stage that the average engagements of local clips is that of the video label. To verify the effectiveness of our method, we compare the performance of the proposed approach with that of several state-of-the-art peer solutions through extensive experiments.

1 Introduction

The emergence of Massive Open Online Courses (MOOC) has attracted wide attention and great expectation from the broad education community. Despite the promising potential of the new education avenue, poor student retention rates are recognized as one of its major caveats. To combat the deficiency, dynamic assessment of individual students’ engagement during their online learning activities can offer just-in-time instructional intervention to improve retention rates and personalized learning outcomes, whose effectiveness has been abundantly supported in the education literature [Dewan

* Jiayao Ma and Xinbo Jiang contribute equally.

† Xueying Qin and Songhua Xu both are correspondence authors.

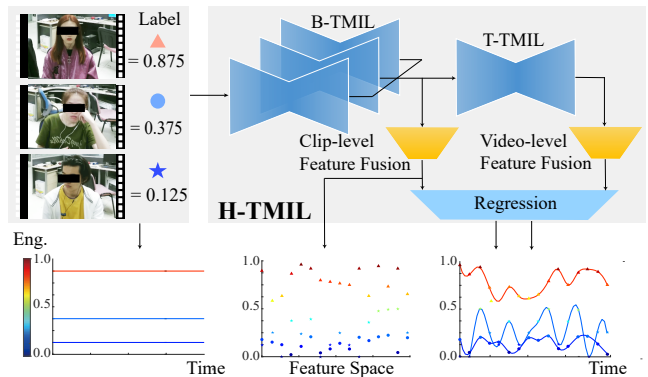


Figure 1: The main idea behind the proposed automatic learning engagement assessment pipeline.

et al., 2019; Zeng *et al.*, 2020]. Due to the large number of students frequently seen in a MOOC environment, manually conducting such assessment is prohibitively expensive. Therefore, it is strongly desirable to develop automated technologies capable of assessing student learning engagement on the fly.

The research on the automatic assessment of learning engagement has the following problems: 1) Due to the time-consuming and laborious work of annotating clip by clip, most previous methods [Osokin, 2019; Wang *et al.*, 2019] only solve the problems of assessing the learning engagement of the whole video, although it makes more sense to assess each short clip. 2) A course in distance education usually lasts for tens of minutes or even an hour, making it difficult for assessment. How to obtain effective features that can represent the entire video and each short clip has become an urgent problem to be solved.

In response to the problems mentioned above, we construct a novel hierarchical multiple instance learning (MIL) engagement assessment model as shown in Fig. 1. Based on the features extracted from video frames, a hierarchical neural network that consists of a clip-level MIL module and a video-level MIL module is trained with only video-level labels. The network can order short clip-level features with their engagements, which is achieved by constraining on a global loss for the top module and a local loss for the bottom module, to ensure that the trained network can accurately and reliably

infer local clip-level labels for a collection of temporally-sequenced clip-level instances under the supervision of global video-level labels. In summary, the main contribution of this paper are threefold as follows:

- We propose a novel temporal multiple instance learning framework which is trained with video-level labels only but can both assess the learning engagement of the full-length video and its constituent short clips.
- We construct a hierarchical neural network which comprises a bottom module and a top module with their tailor-designed loss functions to learn the latent relationship in the frame-clip-video structure.
- A new data set about online course studies is collected and some ablation experiments and comparative experiments are conducted on it to prove that the proposed method can effectively assess the learning engagement of both the whole video and each video clip.

2 Related Work

Learning Engagement Assessment Based on Visual Clues

When a learner is studying, the camera can easily capture the learner’s images and advanced features associated with learning engagement can be effectively extract from these video frames. Some methods assess learning engagement based on RGB videos due to their easy accessibility. Studies [Whitehill and Movellan, 2008; Whitehill *et al.*, 2014] have found that facial expressions and head poses are directly related to learning engagement. [Frank *et al.*, 2016] proposes a framework for engagement assessment, which uses facial expressions, sound, body poses and movements acquired through multiple sensors, and then a SVM classifier is used to classify engagement. [Chang *et al.*, 2018; Yang *et al.*, 2018] uses features such as facial expressions, eye parameters, gestures and head poses extracted from image sequences to recognize the learner’s cognitive state. Our work only uses body poses, head poses and face landmarks as the input features.

Multiple Instance Learning The multiple instance learning(MIL) method is a solution to the problem of weakly supervised learning, which trains models based on weakly labeled data. The MIL is widely used in medical image diagnosis and detection [Li *et al.*, 2019; Campanella *et al.*, 2019; Rony *et al.*, 2019]. In this problem, the training set consists of bags labeled as 0 or 1, and each bag contains many instances without labels. Taking cancer prediction as an example here, each image with a cancer/non-cancer label forms a "bag", and each pixel or block of pixels that constitutes the image is called an "instance". We divide the MIL works into three categories: the global MIL, which can determine whether there is a disease [Campanella *et al.*, 2019; Dov *et al.*, 2019]; the local MIL, which can mark diseased cells on the image [Xu *et al.*, 2019; Liang *et al.*, 2018]; the global and local MIL, which can detect whether there is a disease and indicate the diseased cells at the same time [Wang *et al.*, 2020; Huang and Chung, 2019]. Compared with the above problems, learning engagement assessment is the same that the

bag composed of multiple instances has a label, but each instance has no label. But the difference is that there are structural and temporal correlations between the instances, and the label of the bag is not limited to 0 or 1 but any real number within the label range.

3 Method

3.1 Problem Formulation

In this video-based learning engagement assessment task, we first divide the video evenly into N clips, and each clip \mathbf{X}_i is composed of l video frames, so the whole video can be represented as its sample clip sequence $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$, where the video is a bag and its sample clip sequence is a instance sequence. Furthermore, each clip \mathbf{X}_i can be represented as $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,l}\}$ and $\mathbf{x}_{i,j}$ is the j th frame of the i th clip, where the clip is a bag and its frames are instances. Based on this frame-clip-video structure, we use the hierarchical multiple instance learning module from bottom to top to gradually use frame-level features to aggregate the clip-level descriptors and then aggregate the video-level descriptor based on them. We first use the model g to extract the feature sequence $\mathbf{F}_i = \{\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \dots, \mathbf{f}_{i,l}\}$ from the frames of i th video clip, where $\mathbf{f}_{i,j} = g(\mathbf{x}_{i,j})$. Then the bottom module D_b uses \mathbf{F}_i as the input to learn the clip-level feature \mathbf{C}_i . The clip-level feature sequence $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_N\}$ is used as the input of top module D_t to learn the video-level feature \mathbf{V} . The video-level feature \mathbf{V} and the clip-level feature \mathbf{C}_i are respectively sent to the regression function Reg to obtain the video-level learning engagement \hat{Y} and the clip-level learning engagement sequence $\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_N\}$. The process can be modeled as follows:

$$\hat{Y} = Reg(D_t(D_b(g(\mathbf{X}), \varphi), \theta), \mu) \quad (1)$$

and

$$\tilde{y}_i = Reg(D_b(g(\mathbf{X}_i), \varphi), \mu) \quad (2)$$

where φ, θ, μ are trainable parameters. The only supervision information we have is the video-level label Y , and Y is a decimal between 0 and 1. The framework is shown in Fig. 2.

3.2 Preprocessing

Down-sampling Learners’ head poses and facial landmarks tend to change gradually and slowly during his/her online studies. Therefore, we down-sample every raw video by keeping one frame every few frames for more efficient computational processing. In our experiments, 3000 frames for each video are reserved for assessment.

Segmentation Because the feature has little effect on the assessment of learning engagement at the time far away from it, and the network is usually difficult to deal with lengthy feature sequences, we segment an input video into short video clips as the basic analysis objects. We set the length of a video clip to be 30 frames, denoted as $l = 30$ in all our experiments to yield an empirically optimized trade-off between computational efficiency and accuracy of the proposed approach. The number N of video clips extracted from each video is 100.

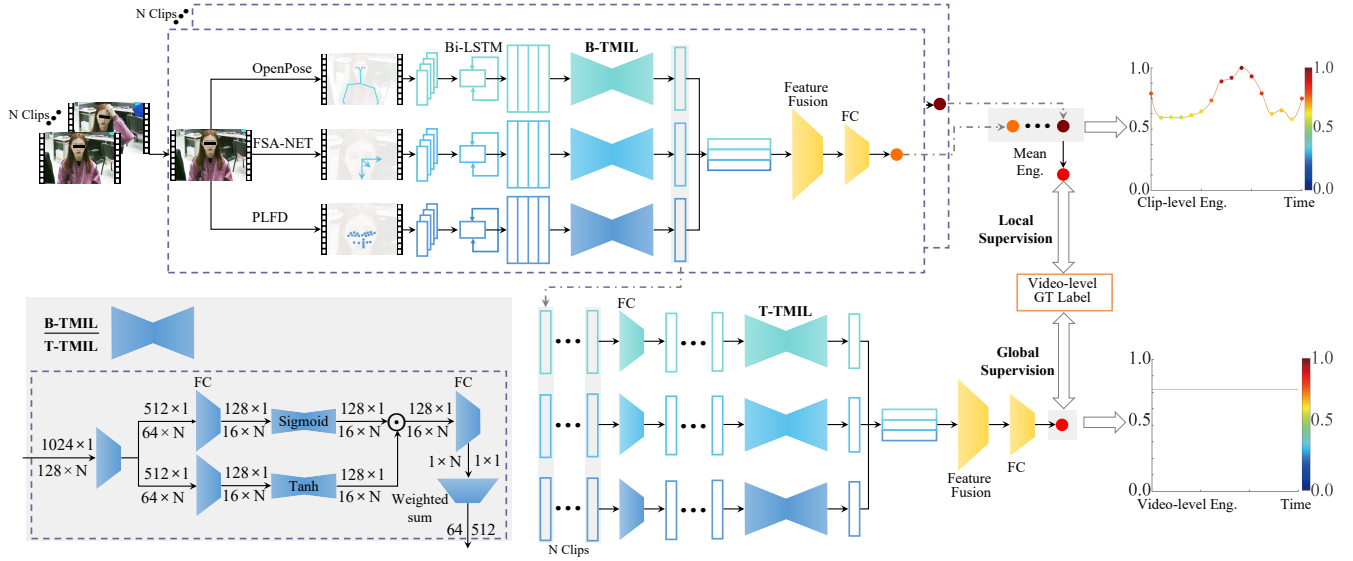


Figure 2: The framework of the proposed network. The main structure of the proposed network are both the bottom temporal multi-instance learning(B-TMIL) and the top temporal multi-instance learning(T-TMIL), which are used to assess video clips and the entire video separately. In the lower left corner of the figure, we give the specific structure of B-TMIL and T-TMIL.

Feature Extraction According to the previous works [Wang *et al.*, 2019; Wu *et al.*, 2019], body languages and facial expressions have a strong correlation with a learner’s learning engagement. Therefore we use head pose features, body pose features and facial landmarks as the input to improve the accuracy and robustness of our model. For video frame $\mathbf{x}_{i,j}$, we use methods proposed by [Yang *et al.*, 2019; Guo *et al.*, 2019; Osokin, 2019] to extract head pose features $\mathbf{e}_{i,j}$, body pose features $\mathbf{b}_{i,j}$ and facial landmarks $\mathbf{m}_{i,j}$ respectively. Then for video clip \mathbf{X}_i , we can get head pose sequence $\mathbf{E}_i = \{\mathbf{e}_{i,1}, \mathbf{e}_{i,2}, \dots, \mathbf{e}_{i,l}\}$, body pose sequence $\mathbf{B}_i = \{\mathbf{b}_{i,1}, \mathbf{b}_{i,2}, \dots, \mathbf{b}_{i,l}\}$ and facial landmark sequence $\mathbf{M}_i = \{\mathbf{m}_{i,1}, \mathbf{m}_{i,2}, \dots, \mathbf{m}_{i,l}\}$.

3.3 Hierarchical Temporal Multiple Instance Learning Module (H-TMIL)

Based on the frame-clip-video structure only with video-level labels, we propose the hierarchical temporal multi-instance learning framework composed of a bottom temporal multi-instance learning module (B-TMIL) and a top temporal multi-instance learning module (T-TMIL) which respectively dedicated to learning the latent relationship between a clip and its constituent frames and that between a video and its constituent clips. Through this framework, we build the connection between the underlying video frames and the video-level label, and can implicitly learn the expression of the middleware, namely the clip-level features which are useful for assessing clip-level learning engagement.

Bottom Temporal Multiple Instance Learning (B-TMIL)

The B-TMIL module acts on sampled video frame sequence which composes a short video clip, where the frames are instances and the clip is the bag. We need to obtain the bag’s valid representation in order to accurately obtain its label. However, unlike the traditional MIL instances, for a

short-time frame sequence, there is a strong timing correlation between frames. Some methods [Wang *et al.*, 2019; Huynh *et al.*, 2019; Wu *et al.*, 2019] use LSTM to capture the temporal association and use the hidden state of the last layer to express the sequence, which will cause the loss of early information. In response to this problem, we use a self-attention multi-instance module to act on all hidden layer features of the bidirectional LSTM built on instances, and adaptively obtain the representation of the bag through trainable parameters. The body pose feature is taken as an example.

First, we input the body pose feature sequence \mathbf{B}_i into the Bi-LSTM to obtain the hidden state sequence $\mathbf{H}_i = \{\mathbf{h}_{i,1}, \mathbf{h}_{i,2}, \dots, \mathbf{h}_{i,l}\}$. Inspired by the paper [Ilse *et al.*, 2018], the clip-level aggregated feature $\tilde{\mathbf{B}}_i$ can be computed as a weighted sum of $\mathbf{h}_{i,j}$ after dimension reduction:

$$\tilde{\mathbf{B}}_i = \sum_{j=1}^l \tilde{\alpha}_{i,j} \delta(\tilde{\mathbf{W}}_1 \mathbf{h}_{i,j}^\top) \quad (3)$$

Here δ refers to the ReLU function and $\tilde{\mathbf{W}}_1$ refers to the full connection operation which are used to reduce dimension. The weight $\tilde{\alpha}_{i,j}$ is computed by:

$$\tilde{\alpha}_{i,j} = \frac{score_{i,j}}{\sum_{o=1}^l score_{i,o}} \quad (4)$$

where $score_{i,j}$ is calculated as followed:

$$score_{i,j} = \tilde{\mathbf{W}}_4 (\sigma(\tilde{\mathbf{W}}_2 \delta(\tilde{\mathbf{W}}_1 \mathbf{h}_{i,j}^\top)) \odot \tau(\tilde{\mathbf{W}}_3 \delta(\tilde{\mathbf{W}}_1 \mathbf{h}_{i,j}^\top))) \quad (5)$$

where $\tilde{\mathbf{W}}_1, \tilde{\mathbf{W}}_2, \tilde{\mathbf{W}}_3, \tilde{\mathbf{W}}_4$ are the weight matrices, \odot is an element-wise multiplication, σ is the Sigmoid function and τ is the Tanh function. Among them, Tanh function is used to obtain the correlation between features, and Sigmoid function acts as a gate mechanism. Similar to the aggregation process

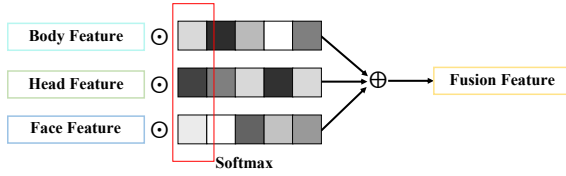


Figure 3: The process of the feature fusion.

of clip-level body pose feature $\tilde{\mathbf{B}}_i$, we use the B-TMIL module to extract clip-level head pose feature sequence $\tilde{\mathbf{E}}_i$ and clip-level facial landmark feature sequence $\tilde{\mathbf{M}}_i$.

Top Temporal Multiple Instance Learning(T-TMIL)

Similar to B-TMIL, T-TMIL acts between video clips which can be thought as instances and the full-length video composed of clips is the bag. However, due to the long duration of each video clip, we believe that there is no longer a strong temporal relationship between video clips. Previous work [Wang *et al.*, 2019; Wu *et al.*, 2019] generally believes that the video-level feature can be expressed as the average of all clip-level features. These methods treat all video clips equally, which adversely affects the final assessment accuracy. In order to get more robust and flexible representation of video, we still apply MIL module on the basis of clip-level features. And the top-level module becomes more similar to the traditional multi-instance structure. Here we also use body pose features as an example to describe the process of the T-TMIL.

The dimension of the clip-level feature is reduced through the fully connected operation to generate more efficient embedding representation. We construct the weighted combination of video clips to represent a video which can be calculated as follows:

$$\hat{\mathbf{B}} = \sum_{i=1}^N \hat{\alpha}_i \delta \left(\hat{\mathbf{W}}_1 \tilde{\mathbf{B}}_i^T \right) \quad (6)$$

$$\hat{\alpha}_i = \frac{score_i}{\sum_{q=1}^N score_q} \quad (7)$$

where $score_i$ is calculated as followed:

$$score_i = \hat{\mathbf{W}}_4 \left(\sigma \left(\hat{\mathbf{W}}_2 \delta \left(\hat{\mathbf{W}}_1 \tilde{\mathbf{B}}_i^T \right) \right) \odot \tau \left(\hat{\mathbf{W}}_3 \delta \left(\hat{\mathbf{W}}_1 \tilde{\mathbf{B}}_i^T \right) \right) \right) \quad (8)$$

where $\hat{\mathbf{W}}_1, \hat{\mathbf{W}}_2, \hat{\mathbf{W}}_3, \hat{\mathbf{W}}_4$ are the weight matrices of T-TMIL. Similar to the above process, we use the T-TMIL module to aggregate video-level head pose feature sequence $\hat{\mathbf{E}}$ and video-level facial landmark feature sequence $\hat{\mathbf{M}}$.

3.4 Feature Fusion

We use three types of features to assess learning engagement, while our hierarchical module processes each type of feature separately. In order to achieve complementary advantages between different features and increase judgment information, we propose a weighted feature fusion method as shown in Fig. 3 and then obtain the clip-level and video-level fusion features for the assessment. The weight matrix has the

same size as the feature matrix composed of the three features and is composed of trainable parameters. Each column of the weight matrix is first normalized by Softmax function to obtain different proportions of different features in this dimension, and then weighted summed to obtain the weighted fusion feature.

3.5 Loss Function

In order to train and optimize the proposed framework, we design two loss functions: the global and the local supervision loss function. The global supervision loss function uses the mean square error method to measure the difference between video-level learning engagement result and the ground-truth. The local supervision loss function also uses the mean square error method but measures the difference between the mean of clip-level learning engagement result and the label of the video. Through this joint loss function and our hierarchical temporal multi-instance learning module, we can constrain both video-level embedding features and clip-level embedding features. The formula of loss is as follows:

$$L_{total} = \alpha L_{local} + \beta L_{global} \quad (9)$$

where α and β are two hyper-parameters. Each component of the loss can be formulated as follows:

$$L_{local} = \frac{1}{K} \sum_{k=1}^K \left(Y_k - \frac{1}{N} \sum_{i=1}^N \tilde{y}_{ki} \right)^2 \quad (10)$$

and

$$L_{global} = \frac{1}{K} \sum_{k=1}^K \left(Y_k - \hat{Y}_k \right)^2 \quad (11)$$

where \tilde{y}_{ki} is the clip-level assessment result of i -th clip of k -th video, \hat{Y}_k is the video-level assessment result of k -th video, and Y_k is the video-level label of k -th video.

4 Experiments

Due to the lack of annotated long video data sets publicly available for our study on learning engagement assessment, we collect and annotate one in-house data set about online course studies. Using the data set, we explore the performance of the proposed network against that of multiple state-of-the-art methods.

4.1 Online Course Studies (OCS) Data set

This data set carries 236 videos captured of 59 subjects while they are watching four healthcare topic courses. In order to obtain more diverse data for the learning engagement, we have carefully selected courses on four topics: cancer prevention, respiratory diseases, renal system and sleep problems. The length of each course is about 30 minutes, and all subjects are required to watch the entire content of each course. Except for a 14-year-old junior high school student, all subjects are undergraduate or graduate students, aged between 20 and 32, majoring in software engineering and digital media. During the learning period, the performances of all subjects are according to personal preferences and status. We will not

Score	0	0.125	0.25	0.375	0.5
Video#(OCS)	22	18	21	24	12
Score	0.625	0.75	0.875	1	
Video#(OCS)	45	53	11	30	
Total	236				

Table 1: Distributions of scores in the OCS Data set.

place any restrictions on the postures and behaviors of subjects.

Each learning process of the subjects will be given a score from 0 to 1 to indicate the level of the learning engagement. The larger the score, the higher the engagement. The score of the engagement is based on three factors: the subject’s self-scoring, the annotator’s score and the testing score. First, we require each subject to self-score according to the actual learning state of himself. Secondly, the annotator is required to carefully watch each learning video and gives the score of the engagement. We invite three annotators to score the same video from 0 to 8 at the same time. If the difference between any two labels is less than 2, the final label is the average of the three labels. Otherwise, the three labelers negotiate and determine the final label. Thirdly, each subject is asked to finish all the questions about the courses. Regarding whether to use test scores, we will make a decision based on the relationship between score distribution and engagement. All three scores will be normalized between 0 and 1. Then the arithmetic average of the three scores is the score of the engagement. Table. 1 shows the distribution of our data set.

In addition, we have obtained the consent of the subjects and can use the collected data for the research of this project. And we will not disclose anyone’s identity information, and everyone’s eye position will be covered with a black mask to protect his personal privacy.

4.2 Implementation Details

Data Processing Based on the trained model, we first obtain head pose, body key points and facial key points from original videos. As shown in Fig. 2, the head pose is composed of yaw, roll and pitch angles and the coordinates of the center point of head (5D vector). Only the upper body of each subject appears in the video, thus the body pose features only need to be composed of the upper body key points(24D vector). The face landmarks are 23 key points (46D vector) extracted from the original face model with 98 key points. The models for feature extraction are light-weight and unstable, whose outputs are sometimes missing. Therefore we perform linear interpolation over $N(N \leq 15)$, assumed that only the missing data within 0.5s is caused by the instability of the model) adjacent vectors with missing values. Then we evenly extract 3000 frames from the final data as input to the model. In addition, in order to eliminate the influence of the subjects at different positions in the screen to the model, we use the coordinates of the neck as the center point for the body pose, and subtract the coordinates of the center point from the coordinates of the remaining points. And for face landmarks, we use the mean value of the coordinates of the key points on the nose as the center point and also subtract the coordinates

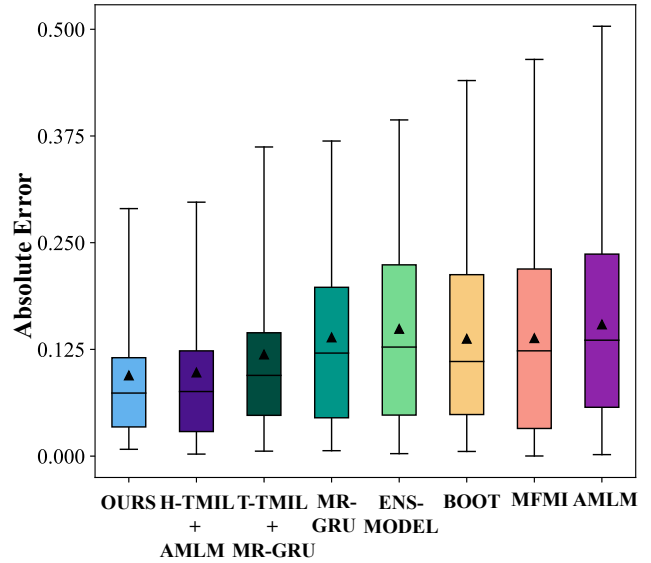


Figure 4: Performance comparison among our and state-of-the-art peer methods.

of the center point from other points. In addition, we found that the number of different types of data in our data set is quite different. To make the distribution of our data set more uniform, we randomly add noise to the features and resample the videos whose type is in a small number in different ways. And we randomly select three quarters of videos in the OCS data set as our train set and the remaining quarter as test set.

Training Setting We establish the project based on Pytorch and train it on Tesla K80G GPU. Our training epoch is set to 200. We initialize the learning rate to 0.001, and use Adam with a momentum of 0.9 and a weight decay of $1e-4$ as the optimizer. When the training epoch is 60/100/160, we multiply the learning rate by 0.1.

Speed Our assessment model’s speed is about 0.014s per video clip and the feature extraction models whose speeds are at least 20 FPS can be done in parallel.

4.3 Ablation Studies

TMIL Module We explore the effectiveness of TMIL module by replacing the TMIL module used for aggregating frame-level and clip-level features with the methods proposed by the previous work [Wu *et al.*, 2019], such as the last hidden state vector of LSTM and simply mean pooling. The experimental results are shown in the Table. 3. When processing frame-level features, only using the last hidden state vector of LSTM performs the worst, which shows that only the final result of LSTM cannot effectively aggregates enough and effective information; as for the result using mean pooling is worse than which using the B-TMIL module, we believe that it is due to the use of self-attention aggregation which enables the model to adaptively find more important features and solve the problem that LSTM can’t deal with long-range dependence effectively. In addition, when using B-TMIL module only for frame-level features, it performs better than that when using T-TMIL module only for clip-level features. This











Frame					
Face	0.0846	0.1185	0.1282	0.1593	0.0879
Head	0.0618	0.0733	0.0681	0.0643	0.0590
Pose	0.0002	0.0003	0.0011	0.0116	0.0943
Frame					
Face	0.1053	0.0501	0.0459	0.0964	0.1239
Head	0.0609	0.1970	0.3195	0.0388	0.0572
Pose	0.2656	0.3256	0.2273	0.0362	0.0379

Table 2: The visualization of the B-TMIL attention of face landmarks, head pose and body pose features on each frame.

Method	MSE	MAE	STD
h_{30} +Mean	0.0241	0.1212	0.0968
Mean+Mean	0.0238	0.1154	0.1023
Mean+T-TMIL	0.0200	0.1066	0.0928
B-TMIL+Mean	0.0189	0.0977	0.0968
B-TMIL+T-TMIL	0.0158	0.0944	0.0830

Table 3: Different methods to aggregate features (A+B means using method A on frame-level features and using method B on clip-level features).

can be easily explained as that model can be more effective only when the basic features, that is, clip-level features are accurately aggregated. Based on this, we obtain more valid video-level features. To further prove the effectiveness of our proposed module, we choose some key frames and visualize the results of the B-TMIL module acting on the underlying video frames on the test set. Table. 2 shows the visualization results of our module on the three features. Since the subjects' facial expression has not changed much, the attention weights are almost in the same order of magnitude. As for the head pose and body pose features, the weights of attention increase significantly (as shown in the bold part in the table) when subject's head and body pose changes. It can be seen that our attention module has effectively paid attention to the parts that are important for obtaining results.

Local and Global Supervision The assessment results of the learning engagement from the entire video can be obtained from a predictor that takes the characteristics of the entire video as input, or the result of each video clip can be obtained first and be averaged as the final result. In the model, in addition to using the loss function for the final video-level result to achieve global supervision, we also add local supervision to constrain the mean of the clip-level results. The experimental results are shown in the Table. 4. We obtain the best results with using global supervision and local supervision at the same time. The result of using global supervision only is not as good as the result of using local supervision

Method	MSE	MAE	STD
Only global supervision	0.0234	0.1167	0.0989
Only local supervision	0.0199	0.1089	0.0896
Global + Local Supervision	0.0158	0.0944	0.0830

Table 4: Effectiveness of local and global supervision

only. This is consistent with our previous description that more effective advanced features can be obtained only when effective basic features are obtained.

4.4 Comparison with State-of-the-Art

To explore the advantages of the proposed approaches in the engagement assessment task, we conduct a series of experiments where benchmarked performance of the new approach is compared with that of several peer methods on the OCS, such as **BOOT** [Wang *et al.*, 2019], **ENS-MODEL** [Huynh *et al.*, 2019], **MFMI** [Wu *et al.*, 2019], **MR-GRU** [Zhu *et al.*, 2020] and **AMLM** [Wu *et al.*, 2020]. In addition, we also optimize MR-GRU and AMLM with our TMIL module. Since there is already an attention mechanism in MR-GRU, we only add the T-TMIL module to it, while for AMLM, we add the complete H-TMIL module. All peer methods are reimplemented due to the lack of open source alternatives.

For the task of assessing video-level learning engagement, Table. 5 shows the video-level mean absolute error (MAE) and mean squared error (MSE) on OCS. The original MR-GRU is better than the original AMLM, but the AMLM with our complete H-TMIL module has better performance than the MR-GRU with only the T-TMIL module, which further proves the effectiveness of our proposed module and the importance of aggregating effective underlying features. Fig. 4 shows the absolute error (AE) between algorithmically generated results and their corresponding labels using boxplots.

One of our most important contributions is that the proposed network can assess the learning engagement of video clips relatively accurately only using video-level labels. In order to demonstrate the effectiveness of our model on the task of assessing clip-level results, we annotate some videos clip

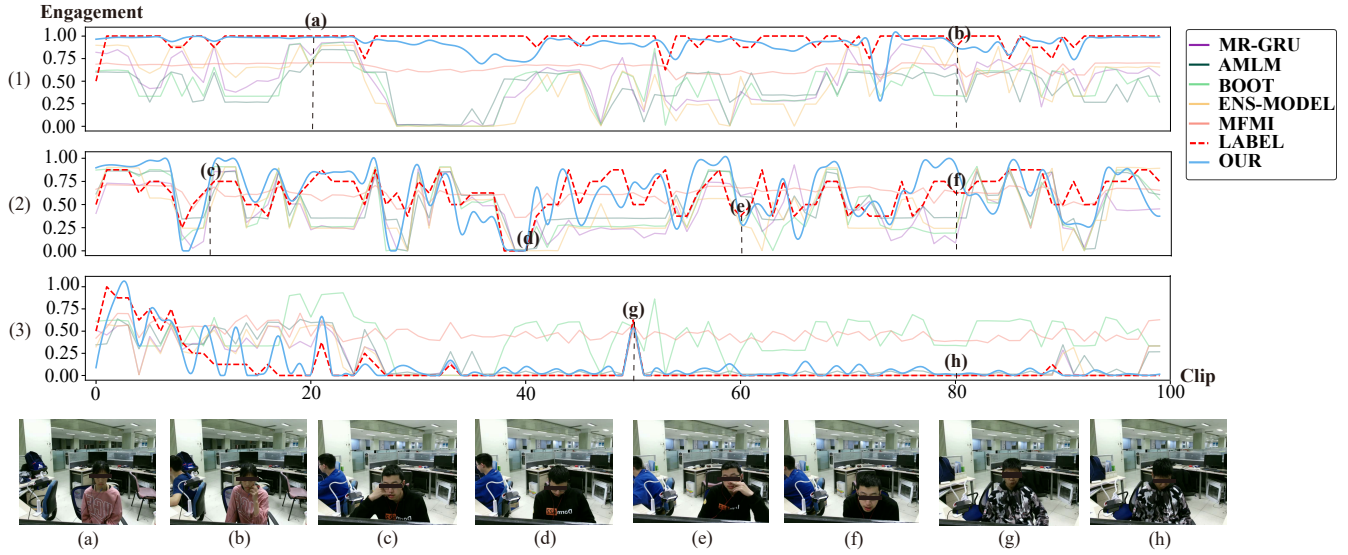


Figure 5: The visualization of clip-level learning engagement of three videos respectively with label 0.875, 0.5 and 0.125. The frame (a) to (h) are some key frames.

Method	MSE	MAE
AMLM	0.037	0.154
ENS-MODEL	0.035	0.149
MFMI	0.032	0.138
MR-GRU	0.030	0.139
BOOT	0.030	0.137
Our T-TMIL+MR-GRU	0.026	0.119
Our H-TMIL+AMLM	0.017	0.098
OURS	0.0158	0.094

Table 5: Video-level results on OCS Data set.

by clip. Fig. 5 (1), (2) and (3) respectively represent the clip-level results and corresponding labels of the five models in three videos with different learning engagement scores. The result curve of our method and the label curve basically maintains the same trend which fully demonstrates that our model can effectively learn clip-level features only with video-level labels provided. Table. 6 further presents the performance of the proposed approach with that of other five peer methods and two models improved with our modules in terms of the mean absolute error (MAE), mean squared error (MSE). Our model achieves remarkable results except on the video (2) where the range of labels has a relatively large float. Although the MFMI model has some advantages in overall statistical results, it does not well reflect the drastic changes of the engagement, as shown in the pink curve of Fig. 5 (2). All these results demonstrate the superiority of our approach compared with all peer methods in assessing clip-level learning engagement.

5 Conclusion

To solve the problem of assessing the changes of learning engagement of students in the learning process with only video-level labels, a novel hierarchical temporal multi-instance learning network is introduced to automatically assess both

Method	Video (1)	Video (2)	Video (3)
ENS-MODEL	0.333/0.493	0.097/0.248	0.035/0.094
MFMI	0.105/0.314	0.035/0.157	0.186/0.416
BOOT	0.345/0.531	0.094/0.241	0.206/0.388
AMLM	0.333/0.528	0.064/0.197	0.026/0.098
MR-GRU	0.297/0.465	0.107/0.253	0.020/0.077
H-TMIL+AMLM	0.047/0.183	0.062/0.192	0.014/0.092
T-TMIL+MR-GRU	0.110/0.298	0.067/0.191	0.016/0.101
Ours	0.018/0.084	0.051/0.190	0.014/0.069

Table 6: Clip-level results on OCS Data set (MSE/MAE).

video-level and clip-level learning engagement from video streaming data in an end-to-end learning fashion composed of a bottom and a top module which can respectively learn the latent relationship between a video and its constituent clips and that between a clip and its constituent frames.

Future Work Our future work mainly consists of two parts. 1) More effective structure. Unify feature extraction and assessment into an end-to-end model and introduce more advanced structures to get high-level semantic information more effectively. 2) Practicality. Build a system that can assess learning engagement online and prove the versatility of model in other scenarios, such as evaluating the user’s interest of advertisements or products and the degree of fatigue in driving.

Acknowledgements

This work was supported by Zhejiang Lab under Grant(No.2020NB0AB02), NSFC of China under Grant(No.61907026) and Shandong Province Higher Educational Science and Technology Program under Grant(No.J18KA392).

References

[Campanella *et al.*, 2019] Gabriele Campanella, Matthew G Hanna, and et al. Clinical-grade computational pathology

- using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.
- [Chang *et al.*, 2018] Cheng Chang, Cheng Zhang, Lei Chen, and Yang Liu. An ensemble model using face and body tracking for engagement detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 616–622, 2018.
- [Dewan *et al.*, 2019] M Ali Akber Dewan, Mahbub Mureshed, and Fuhua Lin. Engagement detection in online learning: a review. *Smart Learning Environments*, 6(1):1, 2019.
- [Dov *et al.*, 2019] David Dov, Shahar Z Kovalsky, Jonathan Cohen, Danielle Elliott Range, Ricardo Henao, and Lawrence Carin. Thyroid cancer malignancy prediction from whole slide cytopathology images. In *Machine Learning for Healthcare Conference*, pages 553–570. PMLR, 2019.
- [Frank *et al.*, 2016] Maria Frank, Ghassem Tofighi, Haisong Gu, and Renate Fruchter. Engagement detection in meetings. *ArXiv*, abs/1608.08711, 2016.
- [Guo *et al.*, 2019] Xiaojie Guo, Siyuan Li, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. PFLD: A practical facial landmark detector. *CoRR*, abs/1902.10859, 2019.
- [Huang and Chung, 2019] Yongxiang Huang and Albert CS Chung. Evidence localization for pathology images using weakly supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 613–621. Springer, 2019.
- [Huynh *et al.*, 2019] Van Thong Huynh, Soo-Hyung Kim, Gueesang Lee, and Hyung-Jeong Yang. Engagement intensity prediction with facial behavior features. In *International Conference on Multimodal Interaction, ICMI*, 2019.
- [Ilse *et al.*, 2018] Maximilian Ilse, Jakob M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.
- [Li *et al.*, 2019] Jiachen Li, Songhua Xu, and Xueying Qin. Cervical vertebrae health score method based on multiple instance learning. *Journal of Computer-Aided Design Computer Graphics*, 31(01):96–105, 2019.
- [Liang *et al.*, 2018] Qiaokang Liang, Yang Nan, Gianmarco Coppola, Kunlin Zou, Wei Sun, Dan Zhang, Yaonan Wang, and Guanzhen Yu. Weakly supervised biomedical image segmentation by reiterative learning. *IEEE Journal of biomedical and health informatics*, 23(3):1205–1214, 2018.
- [Osokin, 2019] Daniil Osokin. Real-time 2d multi-person pose estimation on CPU: lightweight openpose. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2019, Prague, Czech Republic, February 19-21, 2019*, 2019.
- [Rony *et al.*, 2019] Jérôme Rony, Soufiane Belharbi, Jose Dolz, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Deep weakly-supervised learning methods for classification and localization in histology images: a survey. *arXiv preprint arXiv:1909.03354*, 2019.
- [Wang *et al.*, 2019] Kai Wang, Jianfei Yang, Da Guo, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. Bootstrap model ensemble and rank loss for engagement intensity regression. In *International Conference on Multimodal Interaction, ICMI*, 2019.
- [Wang *et al.*, 2020] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, M. Cai, and P. Heng. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50:3950–3962, 2020.
- [Whitehill and Movellan, 2008] Jacob Whitehill and Javier Movellan. Automatic facial expression recognition for intelligent tutoring systems. *Proceedings of the CVPR Workshop on Human Communicative Behavior Analysis*, 2008.
- [Whitehill *et al.*, 2014] Jacob Whitehill, Zewe Serpell, Yi-Ching Lin, Aysha Foster, and Javier Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *Affective Computing, IEEE Transactions on*, 2014.
- [Wu *et al.*, 2019] Jianming Wu, Zhiguang Zhou, Yanan Wang, Yi Li, Xin Xu, and Yusuke Uchida. Multi-feature and multi-instance learning with anti-overfitting strategy for engagement intensity prediction. In *International Conference on Multimodal Interaction, ICMI*, 2019.
- [Wu *et al.*, 2020] Jianming Wu, Bo Yang, Yanan Wang, and Gen Hattori. Advanced multi-instance learning method with multi-features engineering and conservative optimization for engagement intensity prediction. pages 777–783, 2020.
- [Xu *et al.*, 2019] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10682–10691, 2019.
- [Yang *et al.*, 2018] Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pages 594–598, 2018.
- [Yang *et al.*, 2019] Tsun Yi Yang, Yi Ting Chen, Yen Yu Lin, and Yung Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [Zeng *et al.*, 2020] Haipeng Zeng, Xinhuan Shu, Yanbang Wang, Yong Wang, Liguozhang, Ting-Chuen Pong, and Huamin Qu. Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE Transactions on Visualization and Computer Graphics*, 2020.
- [Zhu *et al.*, 2020] Bin Zhu, Xinjie Lan, Xin Guo, Kenneth E. Barner, and Charles Bonchelet. Multi-rate attention based gru model for engagement prediction. pages 841–848, 2020.