Published in final edited form as:
Int Symp Mix Augment Real. 2020 November ; 2020: 672–681. doi:10.1109/ismar50242.2020.00097.

An Intelligent Augmented Reality Training Framework for Neonatal Endotracheal Intubation

Shang Zhao[1,*], **Xiao Xiao**[1,*], **Qiyue Wang**[1,*], **Xiaoke Zhang**[1,†], **Wei Li**[1,*], **Lamia Soghier**[2,‡], **James Hahn**[1,§]

[1]George Washington University

[2]Children's National Health Systems

Abstract

Neonatal Endotracheal Intubation (ETI) is a critical resuscitation skill that requires tremendous practice of trainees before clinical exposure. However, current manikin-based training regimen is ineffective in providing satisfactory real-time procedural guidance for accurate assessment due to the lack of see-through visualization within the manikin. The training efficiency is further reduced by the limited availability of expert instructors, which inevitably results in a long learning curve for trainees. To this end, we propose an intelligent Augmented Reality (AR) training framework that provides trainees with a complete visualization of the ETI procedure for real-time guidance and assessment. Specifically, the proposed framework is capable of capturing the motions of the laryngoscope and the manikin and offer 3D see-through visualization rendered to the head-mounted display (HMD). Furthermore, an attention-based Convolutional Neural Network (CNN) model is developed to automatically assess the ETI performance from the captured motions as well as identify regions of motions that significantly contribute to the performance evaluation. Lastly, augmented user-friendly feedback is delivered with interpretable results with the ETI scoring rubric through the color-coded motion trajectory that classifies highlighted regions that need more practice. The classification accuracy of our machine learning model is 84.6%.

Index Terms:

Computing methodologies—Computer graphics—Graphics systems and interfaces—Mixed / augmented reality; Computing methodologies—Modeling and simulation—Simulation types and techniques—Real-time simulation; Computing methodologies—Machine learning—Learning paradigms—Supervised learning; Human-centered computing—Visualization—Visualization techniques—Heat maps

1 Introduction

Neonatal endotracheal intubation (ETI) is an essential resuscitation skill for the ventilation of the newborns [20, 25]. Mastering such skill is often complicated by narrow airways, relatively larger tongues compared to adults, anterior glottic positions, and low respiratory

[*] edwinz@gwmail.gwu.edu, xxfall2012@gwmail.gwu.edu, wangqiyue@gwmail.gwu.edu, gw_liwei@gwmail.gwu.edu. [†] xkzhang@gwu.edu. [‡] lsoghier@childrensnational.org. [§] hahn@gwu.edu.

reserves of neonates. The Neonatal Resuscitation Program (NRP) training curriculum requires healthcare providers to undertake tremendous practice to master the ETI procedure with the goal of successfully completing an ETI procedure within 30 seconds [38]. Current ETI practice is often conducted on high-fidelity manikin simulators with an instructor monitoring the practice and providing feedback to trainees [40]. However, current manikins are not capable of providing sufficient procedural information for both instructors and trainees due to the lack of internal situational awareness within the manikin [39,45].

Moreover, trainees have limited understanding of the practice trials they perform because it is difficult to identify undesirable movements merely based on trial outcomes. Hence ETI trainings are heavily reliant on instructors' feedback and evaluations. However, the availability of instructors is often restricted by their substantial clinical duties, which hinders trainees from obtaining sufficient training opportunities. Therefore, it is essential to develop a new training paradigm to accelerate skill acquisition of healthcare providers and increase their training opportunities.

In medical training, research has shown that visual feedback enabled by augmented visualization is beneficial for accelerating medical skill acquisition [42,47]. Augmented Reality (AR) based training systems have emerged as a promising solution to provide situational awareness during ETI procedures, such as by visualizing the laryngoscopic view [8,35]. However, most of the visual feedback in AR based training systems only provides the perceptions during the ETI procedure without the automated assessment. This delivers limited constructive summative feedback for skill acquisition due to the difficulty of identifying critical motions [10,47]. Consequently, these AR based training systems still require feedback from instructors, including monitoring and evaluating the procedures. On the other hand, while classic machine learning algorithms have been used to provide assessments with user-friendly interpretable feedback in medical procedure evaluations [15,16,23], these methods have limited capabilities for complex medical procedure evaluations because of the information loss that results from using hand-crafted features [43]. Convolutional Neural Network (CNN) [17,34,43] has shown an impressive representation power of convolutional features in the medical skill assessment but often lacks user-friendly interpretation. Therefore, it is challenging to integrate automated assessments with AR based training systems, providing both reliable performance evaluation and user-friendly feedback. Relatively little work has been done to explore the intelligent ETI training approach that combines the automated assessment with AR simulation.

In this paper, our main contribution is the development of a new intelligent AR training framework for facilitating ETI practice, which integrates AR visualization with an interpretable machine learning model to address critical issues in the current training regimen. Specifically, the technical contributions of this paper are as follows:

- Build a real-time AR simulation system that captures entire motions of the manikin and the laryngoscope for rendering 3D see-through visualization of the ETI procedure and extracting kinematic multi-variate time-series (MTS) data.

- Develop an automated assessment procedure that provides real-time performance evaluation by using an attention-based CNN, which is better at learning salient motion patterns to predict the level of performance.
- Generate user-friendly interpretations of ETI procedures that combine the assessment rubric of ETI training with the localization of contributing regions of the motion for CNN prediction by using the Generalized Gradient-weighted Class Activation Mapping (Grad-CAM++).
- Create augmented visualization tools for assessment interpretations that offer detailed feedback to trainees through the head-mounted display (HMD), including the summative information of ETI performance and the color-coded motion regions that require further improvement.

2 Related Works

2.1 Medical AR Simulation

With the increased demand of healthcare providers, there is a pressing need for innovative training modalities, such as AR, to develop efficient training platforms and facilitate the medical skill acquisition. Most of the modern AR frameworks for medical training rely on high-fidelity physical manikins and the optical see-through HMD (OST-HMD), which combines additional information from the virtual world with objects that reside in the real world [4,36]. To accomplish complex training tasks that involve situational awareness and proper motion trajectory, medical AR systems integrate various techniques such as motion tracking, registration, and visualization [10].

For ETI training, there are only a few AR-based ETI simulators that have been developed in recent years. Alismail et al. [2], Carlson et al. [8], and Matava et al. [33] presented video laryngoscope applications with HMD to help trainees better recognize the glottis by providing 2D video laryngoscopic view on the display. Although these works provide additional visualization for trainees to understand the procedure, they do not offer free viewpoint visualization to help the instructor assess the performance [10] with spatial perception. To provide users with spatial perception, Hamza et al. [24] and Ballas et al. [6] superimposed the 3D virtual anatomical model over physical manikins, delivering real-time visualization to users during the procedure. Although these visualization tools improve situational awareness, there is only limited feedback on improving their ETI performances. Therefore, current AR-based ETI simulators still require supervision and feedback from the instructor. To the best of our knowledge, none of these works have automated feedback to provide both performance and user-friendly interpretation for ETI procedure. In contrast, our AR training framework provides a solution for combining automated feedback with an AR-based simulator, which can reduce the intervention of instructors.

2.2 Feature Extraction for Motion Analysis

Motion features are critical to represent the discriminative characteristics for motion analysis. For ETI training, electromagnetic (EM) [6, 37, 48], inertial [7], and optical sensors [24] have been used to capture the motion of instruments and limited fiducial points on manikins or patients. To acquire accurate force measurement, Garcia et al. [21] applied force transducers and pressure-sensitive films to measure the forces imparted on specific parts of the manikin during the procedure. Besides accurate force measurements, Delson et al. [11] further evaluated the global movement features that contain the limited information of motion characteristics, such as path length. Although experimental results of these works have shown that the hand-crafted features, such as global movement, have representation power to distinguish in simple procedures, it is insufficient to distinguish the performance of a complex medical procedure with only limited information. For example, the global movement features only focus on the most significant information, which causes the loss of discriminative local trajectory information. In addition, the hand-crafted features are also prone to have co-linearity, which results in only a few of them contributing to the analysis. For example, the hand pose adjustments in pitch and yaw direction tend to occur simultaneously which causes co-linearity for performance prediction. These factors lead to insufficient representation power for performance analysis. Instead of using global movement features, we directly use convolutional features learned by CNN from the kinematic MTS data without any human intervention. Also, the kinematic MTS data can maintain the robustness of the input representation without losing any discriminative information.

2.3 Automated Assessment of Medical Procedure

There is an increasing need for developing automated assessment systems that alleviate the increasing shortage of instructors and provide consistent, quantitative, and objective assessments for better training outcomes. It is possible to develop automated assessment systems with machine learning algorithms to emulate the analysis of human instructors in complex medical procedures with discriminative features extracted from motions. Pairwise preference classifier [32] was proposed to simplify the complex procedure assessment. But its results are highly dependent on the chosen reference for comparison. Traditional machine learning algorithms, such as Support Vector Machine, Hidden Markov Models, and linear regression, have been used in performance evaluation and gesture classification for complex medical procedures [1, 49–51]. However, these approaches require feature selection methods to determine the discriminative features for prediction. Therefore, their predictability is limited by the selected features. In contrast, neural network, an emerging technique of machine learning, implicitly realizes the feature selection by emphasizing the discriminative motion patterns directly from captured data. CNN, which is a representative architecture of neural networks, has been applied to solve classification and regression problems with time-series data. Wang et al. [44] proposed several baseline methods for processing MTS data, such as ResNet [26] and Fully Convolutional Network (FCN) [31]. However, the datasets of medical procedures usually have limited sample sizes due to data privacy and the high cost of acquiring data, resulting in poor training results. To address the limited training samples, Wang et al. [43] introduced a data augmentation method to increase the sample size by segmenting the kinematic MTS data into small fragments with a fixed length. In addition, Fawaz et al. [17,18] introduced FCN to realize surgical performance evaluation with kinematic MTS data by applying the Global Average Pooling (GAP) [30] to reduce the complexity of the network. However, these methods cannot guarantee that their networks can make predictions based on the meaningful local motion patterns. Recently, image-based medical skill assessment had been proposed in the computer vision community [13,14,53]. In particular, Doughty et al. [13,14] have demonstrated that local attention can facilitate the representation power of neural networks. Therefore, our CNN model adopts the attention mechanism with a convolutional block attention module (CBAM) [46] which generates a weight map that corresponds to the specific label at each layer. These attention weight maps guide the CNN to focus more on the discriminative local motion patterns to the final prediction. Although machine learning algorithms have been used to assess performance in other medical procedures, there is relatively little research on ETI automated assessment. Moreover, to the best of our knowledge, our approach is the first attempt to integrate the automated assessment of ETI performance with AR simulations.

2.4 Interpretable Methods for CNN

Although CNN achieves impressive performance on various medical tasks, the “black box” effects from layered perceptual operations make it challenging to develop user-friendly interpretation. In the computer vision community, Class Activation Mapping (CAM) [52] evaluates the contributions of each input data element with respect to a specific label by using the weighed combination of feature map of the last convolutional layer with the weight at the GAP layer by backpropagation. Therefore, the heat map of the CAM results provides localization of discriminative regions, which adds visual explanations to the neural network prediction. Later, Gradient-weighted Class Activation Mapping (Grad-CAM) [41] has been proposed to use the gradient of backpropagation at the last convolutional layer to weight the feature map, which has no limitation on network architecture. However, these methods focus more on a single object or pattern in their results [9]. Wang et al. [44] first introduced one-dimensional CAM to address the interpretability problem of time-series data classification. Furthermore, Fawaz et al. [18] first applied one-dimensional CAM results to visualize surgical performance assessment, but the approach lacked user-friendly explanation to facilitate optimal surgical performance. In contrast, we use not only Grad-CAM++ [9] to provide better localization of multiple pattern instances from kinematic MTS data, but also combine them with the ETI scoring rubric to offer a user-friendly explanation for the regions that need more practice. Specifically, we can correlate the movement patterns defined in the ETI scoring rubric with the discriminative regions of the Grad-CAM++ results to achieve further explanation. Therefore, we can provide both performance evaluation and interpretable feedback with only one neural network.

3 Framework Overview

The proposed intelligent AR training system includes a standard fullterm Laerdal® task trainer manikin, a laryngoscope with a Miller 1 blade, a 3.0 mm endotracheal tube, a 3D Guidance® trakStar™ motion tracking system with 3 EM sensors (6 degrees of freedom (DOF)), and a Microsoft® HoloLens™ HMD. The virtual model for the task trainer was developed by CT scanning the manikin. We extracted the segmented mesh from the volumetric data using 3D Slicer4 [19]. A virtual laryngoscope was modeled on measurements of a real Miller 1 blade and a laryngoscope handle. The HMD allows users to experience 3D see-through visualization. AR devices, such as Microsoft® HoloLens™, require remote data streaming and therefore, we developed a distributed framework to stream the motion data from the computer to the HoloLens. Our distributed framework is comprised of calibration module, communication module, assessment module, and visualization module (Fig. 2). The entire framework was implemented in C++ and Python. Camera calibration and marker detection in the calibration module were realized by OpenCV3.4. The visualization modules of the server and the client were implemented with OpenGL and DirectX, respectively. The client visualization on the HMD provides superimposed virtual models and all augmented feedback, which improves trainees' situational awareness during the training. In addition, the server renders the virtual counterparts of instruments on the computer screen, which provides additional visualization for review. In what follows, we will explain each module in detail.

4 Calibration and Communication

Calibration is crucial in capturing accurate motion data for a multimodal system. We used EM tracking system for system calibration (Fig. 3). The calibration module registered all instruments with their virtual counterparts by evaluating transformations of sensors into a unified global coordinate space. We set the EM transmitter base as the origin of the global coordinate. The system calibration contains the following steps: Firstly, we evaluated $T_{model \mapsto sensor}$ matrices between virtual instruments and corresponding attached sensors by using the predefined fiducial markers, including transformations for the laryngoscope and the manikin. To evaluate the transformation between the base space and the AR HMD space, a 3D-printed calibration marker was designed with the Aruco marker [22] and the EM sensor attached at known relative positions. Then the correspondence set of 3D points can be collected by sampling positions of the marker center in the AR HMD space and corresponding positions of the attached EM sensor in the transmitter base space. Finally, the optimal transformation $T_{base \mapsto hmd}$ can be calculated from the correspondence set by using the singular value decomposition algorithm [3]. After the system calibration, both the laryngoscope and the manikin were registered to their virtual counterparts so that we can easily transform between the model space and the AR HMD view space for tracking and visualization with the following equation:

$$T_{model \mapsto view} = T_{hmd \mapsto view} \cdot T_{base \mapsto hmd} \cdot T_{sensor \mapsto base} \cdot T_{model \mapsto sensor} \quad (1)$$

Note that $T_{model \mapsto view}$ and $T_{sensor \mapsto base}$ represent the transformation from virtual model space to the AR HMD view space and the transformation from sensor to the base space, respectively.

For the communication module, we implemented both the server and the client in an asynchronous manner. We applied user datagram protocol (UDP) to achieve real-time transmission rates for streaming different kinds of data packets, such as command data, motion data, assessment data, and visualized feedback data. To maximize the data throughput, we created multiple threads for the server and the client and assigned a unique port number to each communication task.

5 Automated Intelligent Assessment

Although the neural network has been applied to efficiently tackle various complex tasks, its poor interpretability hinders wide use in medical domains. To address this problem, we developed an intelligent assessment module consisting of both attention mechanisms and Grad-CAM++ explanation. As a result, the system provides both real-time evaluations and prediction interpretations for visualization.

5.1 Study Design

We collected an ETI motion dataset that includes 193 trials performed by 45 subjects. The study was approved by the Institutional Review Board of the cooperative institution. Our dataset includes both attending neonatologists and pediatric residents, thus preserving the diversity of expert levels and strategies of neonatal ETI. For the realism of the experiment environment, we set up our system on an infant warmer in the neonatal intensive care unit. Before the experiment, each participant had a chance to practice on the manikin until one successful intubation was achieved. Then each subject performed 3 to 5 trials of ETI procedures and motion data of the laryngoscope and the manikin were recorded. The 3D motions of the trials were subsequently played back on the server computer screen and evaluated by 3 expert raters who have more than 6 years as practicing neonatologists. To minimize the risk of subjective bias, each rater was blinded to every participant's identity, and the order of the playback was randomized. For each trial, each rater gave an integer-valued score, 1 (bad), 2(fair), or 3(good), for the overall performance of the procedure. To evaluate the scoring consistency among expert raters, we evaluated the agreements on scoring with the concordance correlation coefficients. The concordance correlation coefficients for all 3 rater pairs are 0.80, 0.87, and 0.82, which confirms the raters' agreements on scoring. Therefore, we could use the dataset with averaged scores to develop a CNN by using motion data as the input features and overall score as the labels.

5.2 Input Preprocessing

Instead of using hand-crafted features, such as path length and total time, which cannot fully characterize motion, we used kinematic MTS features from the raw motion sequences to train the CNN, such as positions, rotations, and their corresponding first derivatives. These features reflect basic motion information, leading to better generalization for the model input. To prevent from involving a huge number of parameters to be trained in CNN, we opted not to include the motions of both the manikin and the laryngoscope. We instead used the relative transformation between the manikin and the laryngoscope to compress the kinematic features in the head sensor space. We set the maximum length of the input data to 60 seconds because ETI attempts can be highly time-variant.

5.3 Attention-Based Dilated CNN

Our framework used an attention-based dilated CNN (Fig. 4) for real-time performance evaluation and localization of motion regions that need more practice. The proposed network had 5 convolutional modules (1 dilated convolutional layer for feature extraction and 4 attention-based dilated convolutional modules for motion pattern extraction), a GAP layer, and a fully convolutional layer. The input of the network was the kinematic MTS data, and the output was a score label of 3 levels. The feature maps in the hidden layer were organized as $L \times C \times 1$, where L represents the sequence length and $C \in \{8, 16, 32, 64\}$ represents the number of feature channels. However, not all the information on kinematic MTS data is useful. For ETI procedure, trainees are prone to conducting some irrelevant movements that contribute little to skill acquisition, which hinders the localization of discriminative motion patterns for ETI evaluation. To make CNN evaluate performance based on discriminative motion patterns, both dilated convolution and attention mechanism in CNN architecture are applied to our intelligent assessment model for improving the representation power. This expanded receptive fields and integrated attention values that focus on the important features.

The receptive field is one of the key factors that determine the representation power of feature maps in the convolutional layers. With a larger reception field of the convolutional layer, the feature map can be evaluated from a wider coverage of the input data. Increasing the depth of the network will generally expand the receptive field and therefore, the high-level features can be extracted at the top layers of the networks. However, the depth of the network is limited by the size of the training dataset and computational resources due to the increasing size of trainable parameters. Adding pooling layer is an alternative to expand the receptive field but this suffers from the loss of input information due to average pooling or max pooling operations. Instead of increasing the depth of network or adding pooling layers, we applied dilated convolution operation to the CNN model. Dilated convolution can expand the receptive field by scaling the stride of kernel without losing any feature information. Therefore, we can use limited data to train a relatively small network for evaluating ETI performance.

The attention mechanism, which was initially proposed in Natural Language Processing (NLP) domain [5] and deployed in CNN for solving computer vision tasks [27], is another key factor that determines the representation power. The core idea of attention mechanism is to learn a weight vector that can amplify the contribution of highly correlated regions, and thus leading to a learned feature space with more representative patterns. In order to provide meaningful AR augmented feedback such as the localization of undesirable motion patterns, we need not only an acceptable classification accuracy but also a meaningful feature space, both of which can encode various motion patterns in different skill levels. To make CNN focus more on utilizing discriminative motion patterns for prediction, we integrated the attention mechanism into our dilated CNN architecture to improve the performance. Specifically, we realized the attention mechanism with CBAM [46]. For each attention-based dilated convolutional block in our model, the attention map will amplify the hidden layer feature map by conducting an element-wise multiplication. Therefore, the attention maps at each block provide helpful guidance on learning discriminative motion patterns with the classifier during training.

5.4 Interpretable Feedback with Grad-CAM++

The neural network suffers from the opacity of inference, which does not provide interpretable results on its prediction. While CAM results are difficult to interpret directly in a user-friendly manner, it is necessary to explore alternative ways, like indirect methods, to offer user-friendly explanations for a specific problem, such as ETI assessment in our case. Our framework contributes a method to provide user-friendly feedback for ETI training based on the Grad-CAM++ results and ETI scoring rubric.

Unlike Grad-CAM [41], which directly evaluates the weighted combination of gradients, Grad-CAM++ uses a weighting function to weight each gradient component thereby achieving better sensitivity in detecting multiple discriminative patterns. The one-dimensional Grad-CAM++ result is a vector of contributing values. The vector of contributions L can be formulated as the linear combination of weights ${w^s}_c$ and the last convolutional layer feature A^c. For a specific score $s \in [1, 3]$, we have

$$L_t^s = \sum_c^C w_c^s \cdot A_t^c, \tag{2}$$

where $c \in [1, C]$ represents the specific channel in the CNN, and $t \in [1, T]$ represents the timestamp of kinematic MTS data. In Grad-CAM++, the weight value is designed as the weighted sum of the gradients $\frac{\partial y_s}{\partial A_I^c}$ from the backpropagation and corresponding weight α_t^{cs} that is derived from gradients.

$$\alpha_t^{cs} = \frac{\frac{\partial^2 y_s}{\left(\partial A_t^c\right)^2}}{2\frac{\partial^2 y_s}{\left(\partial A_t^c\right)^2} + \sum_t^T A_t^c \frac{\partial^3 y_s}{\left(\partial A_t^c\right)^3}} \tag{3}$$

$$w_c^s = \sum_t^T \alpha_t^{cs} \cdot ReLU\left(\frac{\partial y_s}{\partial A_t^c}\right), \tag{4}$$

where $ReLU(\cdot)$ is the Rectified Linear Unit (ReLU) function that suppresses the negative gradients to 0.

Note that the discriminative regions in the ETI kinematic MTS data cannot be interpreted as the regions for the specific label because we do not know whether these regions have a positive or negative meaning for the specific label. Therefore, the discriminative regions of Grad-CAM++ results cannot directly provide interpretable feedback on improving trainees' performance.

Based on the agreements of expert raters on the scoring rubric, the level of proficiency on motion trajectory is an important factor for performance evaluation. The desirable movements for ETI should be smooth and stable, while the reposition (up and down movements) and rocking movements (side to side movements) of the laryngoscope should be considered undesirable movements. Moreover, we generally assume that the undesirable movements are prone to causing the excessive force on gums and deep insertion of the laryngoscope. Therefore, the ETI scoring rubric considers the desired ETI procedures to minimize the occurrences of unstable movements in the motion trajectory.

Inspired by these analysis, the performance classification with 3 score classes can be considered as different combinations of desirable and undesirable movements in each ETI procedure. Grad-CAM++ utilizes the CNN parameters that learn from the training data so that the discriminative regions of Grad-CAM++ results have statistical meaning in contributing to the prediction. Therefore, it is necessary for trainees to pay more attention to these regions when reviewing their ETI procedures. On the other hand, focusing on all the undesirable movements from the ETI motion trajectory is unnecessary because too many non-discriminative patterns will confuse trainees in recognizing crucial regions that need more practice, slowing the procedure of skill acquisition. Therefore, we developed the user-friendly interpretation that identifies the undesirable movements in the discriminative regions of Grad-CAM++ results by projecting Grad-CAM++ results in the ETI motion trajectory. Note that those discriminative regions with smooth and stable movements will be considered as desirable movements. In contrast, the discriminative regions with high-frequency movements will be labeled as undesirable movements. Based on these interpretations, trainees can focus on their motion trajectories in those local discriminative regions instead of paying attention to the entire trajectory.

6 Visualization

Our augmented visualizations can deliver to trainees with not only real-time see-through visualization for situational awareness during the trial but also visual feedback after each procedure so that trainees can freely examine their performance from any viewpoint. During the ETI procedure, trainees can enable see-through visualization which renders the virtual cross-sectional manikin and the virtual laryngoscope (Fig. 1). Trainees can adjust the laryngoscope based on its spatial relations with the internal geometry of the manikin from the see-through visualization. After each ETI procedure, trainees can examine their motion trajectory at any step of an ETI procedure (Fig. 1) in the motion playback based on the captured motion of both the laryngoscope and the manikin. We also rendered a panel to provide some useful parameters, such as penetration and depth (Fig. 5). Penetration and depth provide indirect information about the trainee's motion trajectory, which helps them avoid applying too much force. To offer more precise instructions, phases of movement were derived according to the natural clinical progression of the procedure (Fig. 5). For each phase, we provided some instructions about the important aspects of the motion trajectory. To provide some general information on assessment, the summative feedback was rendered for trainees, such as performance score and total time. To give feedback on trainee's motion trajectory, the interpretable results of Grad-CAM++ were color-coded and mapped to the 3D motion trajectory (Fig. 1). The color-coded trajectory may generate colors that do not transition smoothly, which causes difficulty of identifying discriminative regions. Taking advantage of temporal coherence of motions, we preformed a 1D convolution with Gaussian kernel to the Grad-CAM++ results to address this problem. With this visualized information, trainees can freely review these motion regions in the motion playbacks from various viewpoints.

The visualization of the assessment feedback provides additional information in the ETI training besides the intubation outcome. For manikin-based training, ventilation is the most important measure of success. The indication of successful ventilation is the inflation of the lung (plastic bags) inside the task trainer manikin. Mastering the medical motor skill requires trainees to perform successful intubation with well-controlled movements. However, most of the existing training simulators only rely on intubation outcomes and cannot deliver appropriate assessment feedback. In contrast, our work allows users to monitor the procedure with real-time evaluation and review movements with highlighted discriminative motion regions.

7 Results

In this section, we conducted both ablation studies and comparison studies to demonstrate the effectiveness of the proposed framework and the contribution of attention mechanism with quantitative and qualitative results. The proposed CNN network was implemented with PyTorch1.5 and Python3.7, which was trained with a NVIDIA GTX 1080Ti GPU. For the training procedure, we used Adam optimizer with multinomial cross-entropy as the objective function [28]. L_2 regularization in the optimization and the batch normalization in attention-based dilated convolutional modules were enabled to prevent overfitting. The learning rate was set at 0.0001, and the number of epochs was 600. The trained model was deployed in the framework by using the LibTorch1.5 C++ library.

To validate the model's effectiveness, several ablation studies were conducted to validate the predictability and demonstrate the meaningful pattern localization of the proposed attention-based dilated CNN in our automated assessment model. All configurations in these experiments used the same kernel size of the first convolutional layer. We used ResNet [26] with 1D convolution as the baseline method. The ablation studies were designed from the following aspects: 1. Importance of dilated convolution; 2. Importance of attention mechanism; 3. Importance of jointly applying attention mechanism and dilated convolution. The training set and testing set were generated by the random partition of the scored dataset. The results on classification accuracy are reported in Table 1. Without dilated convolution, the ablation study of attention mechanism shows that both the baseline and the original CNN method can be improved by 1.9% and 2.9% respectively with the attention mechanism. With dilated convolution, attention mechanism can improve performances of the baseline method and the CNN method by 3.8% and 3.1%, respectively. With the attention mechanism, the ablation study of dilated convolution shows that the dilated convolution can improve performance of the ResNet method and the CNN method by 2.1% and 2.9%, respectively. The results also show that the proposed model outperforms the original CNN model by 5.8%, which indicates that our work can provide good performance evaluation. The improved performance demonstrates that the attention mechanism improves the discriminative power of the classification model.

The contribution of the attention mechanism on improving the classification of each score class was also explored. We computed the confusion matrices (Fig. 6) and receiver operating characteristics (ROC) curves (Fig. 7) for all experimental configurations. The confusion matrices show that CNN models are substantially better in predicting the score classes 1 and 3 and slightly worse in predicting the score class 2 than the ResNet models under the same configuration of the ablation study. This indicates that the convolutional features of CNN models can distinguish the bad and good performances better than ResNet models.

Compared to the original configuration (Fig. 6a and Fig. 6b), which has no dilated convolution and attention mechanism, models with dilated convolution (Fig. 6c and Fig. 6d) had a more evenly distributed prediction on all score classes and an improved prediction on the score class 3. For models with attention mechanism (Fig. 6e and Fig. 6f), the results show that the predictions outperform the ones in the original models. Specifically, the predictions of the attention-based CNN model (Fig. 6f) outperformed the original CNN model (Fig. 6b) by 10.0% and 27.0% on the score class 1 and the score class 3, respectively. These results illustrate that the attention mechanism can guide the classifier to focus more on discriminative motion patterns that distinguish bad and good performances, resulting in better discriminative power of the proposed method. We further explored the classification of score 3 with the kinematic MTS data and find that the motions of score level 3 generally are shorter sequences than the ones in the motions of the other 2 score levels, resulting in fewer numbers of effective features. Therefore, integrating attention mechanism shows better representation power of the discriminative model in the kinematic MTS data with limited length. Moreover, the confusion matrix of the proposed model (Fig. 6h) shows that dilated convolution can improve the prediction on score class 2 from the attention mechanism (Fig. 6f).

For a ROC curve, a larger area under the curve (AUC) value indicates better predictability. Note that the macro-averaged and the micro-averaged AUC values of all approaches were larger than or equal to 0.8 (Fig. 7), indicating that all models in the ablation study had excellent predictability. Specifically, most of the AUC values of each score class in CNN models were above 0.8 except the original CNN method (Fig. 7b), which indicated the good predictability of CNN models for each score class. Both dilated convolution and attention mechanism can facilitate the predictability for CNN methods. The AUC results of the proposed model (Fig. 7h) outperformed all the others. Therefore, we can conclude that the proposed model can accurately and reliably evaluate ETI performance. The ablation study shows that we can apply dilated convolution to expand the receptive field and utilize attention mechanism to improve the performance of CNN for ETI assessment.

In addition, we visualized the contributions of the motion segments to illustrate the impact of the attention mechanism on the final prediction. The CNN model with only dilated convolution was considered as the reference model because it has been demonstrated to have better accuracy distributions in all score classes. We compared the proposed model with the reference model by using visualizations of Grad-CAM++ results of all score classes (Fig. 8). We mapped the results to the 3D positional segments of the laryngoscope handle tip. The results show that the attention mechanism can identify multiple discriminative motion regions. In particular, the number of discriminative regions in the proposed model for score class 1 and score class 2 was significantly larger than the number in the reference model. In particular, the results of the reference model for score class 2 (Fig. 8c) can only identify a small fragment of important movements, which show that the reference model cannot learn the discriminative motion patterns well. In the results of score 3, the proposed model with attention mechanism identified more fine-grained local regions than the reference model. From these results, we conclude that our proposed model with the attention mechanism can guide the CNN to make predictions based on different motion patterns instead of a single pattern as the reference method (Fig. 8 c, and e). Moreover, the discriminative regions from the results of the proposed model had a finer-level localization than the ones from the reference model. These results also show that the proposed user-friendly interpretation method is meaningful with the attention mechanism. Note that discriminative regions had not only rocking and reposition movements (dark red circles in Fig. 8d) but also smooth and stable movements (dark green circle in Fig. 8d). This interpretation matched the movement classification in the ETI scoring rubric.

From the classification perspective, a desirable classifier should have significant differences in the softmax scores to that made by the classifier. In addition, the confidence of the localization has a positive correlation with the softmax score of the network for the specific class [46, 52]. The larger softmax score could result in better localization. From the results (Fig. 8), the softmax scores from our proposed model were generally larger than the ones from the reference model by integrating the attention mechanism. This indicates that the larger number of discriminative regions can lead to a more confident prediction for the evaluation of ETI procedure.

These ablation studies demonstrated the effectiveness of the attention mechanism. In the quantitative evaluations, the confusion matrices show that the attention mechanism improved prediction by facilitating the capability of distinguishing the bad and good performance. ROCs and AUCs show that integrating attention achieved reliable predictability in all classes. In addition, the attention mechanism improved softmax scores of the predicted class, which made less confused prediction and improved the localization of discriminative motion regions.

In addition, we compared the existing methods with the proposed framework on our collected dataset. We replaced convolutional operations in these compared architectures with 1D convolution because most of these works used surgery images as input instead of motion data. The fully connected layer for classification was attached to the end of each network. The compared classification accuracy was reported in Table 2. From the results, our method achieved the best performance. In particular, our method is 11.1% better than the MLP method. We also computed confusion matrices for these methods (the yellow box in Fig. 6). From the results, we observed that the FCN (Fig. 6i) classified the scores 1 and 2 well, but suffered from weak discriminative power for scores 2 and 3. Both TCN and MLP methods (Fig. 6j and Fig. 6k) cannot distinguish score classes 1 and 2 well. In contrast, the LSTM (Fig. 6l) could better classify each score class than the other three methods because it benefited from learning the temporal motion patterns. Finally, our method (Fig. 6h) achieved better prediction of each score class than the LSTM method, which shows that the attention-based dilated CNN can learn better temporal motion patterns to evaluate the intubation procedure.

8 CONCLUSION

In this paper, we proposed a new intelligent AR training framework for neonatal ETI training. The framework supports real-time performance evaluation and post-trial playback with augmented visualization. Therefore, it can not only improve the trainee's situational awareness during the procedure, but also provide automated evaluation and feedback for self-practice. Quantitative and qualitative results of the experiments show that the proposed real-time AR framework has the potential to accelerate the progress of acquiring the ETI skill for trainees and examine the skill proficiency of expert neonatologists. In future work, we will explore with generative models to interpolate expert motions for feedforward motion demonstration, which can further extend our intelligent AR training framework to improve the training efficiency of various medical procedures.

Acknowledgments

Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute Of Child Health Human Development of the National Institutes of Health under Award Number R01HD091179.

References

[1]. Ahmidi N, Tao L, Sefati S, Gao Y, Lea C, Haro BB, Zappella L, Khudanpur S, Vidal R, and Hager GD. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. IEEE Transactions on Biomedical Engineering, 64(9):2025–2041, 2017. [PubMed: 28060703]

[2]. Alismail A, Thomas J, Daher NS, Cohen A, Almutairi W, Terry MH, Huang C, and Tan LD. Augmented reality glasses improve adherence to evidence-based intubation practice. Advances in Medical Education and Practice, 10:279, 2019. [PubMed: 31191075]

[3]. Arun KS, Huang TS, and Blostein SD. Least-squares fitting of two 3-d point sets. IEEE Transactions on Pattern Analysis and Machine Intelligence, (5):698–700, 1987. [PubMed: 21869429]

[4]. Azuma RT. A survey of augmented reality. Presence: Teleoperators & Virtual Environments, 6(4):355–385, 1997.

[5]. Bahdanau D, Cho K, and Bengio Y. Neural machine translation by jointly learning to align and translate. In Bengio Y and LeCun Y, eds., 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015.

[6]. Ballas C, Norfleet J, Reihsen T, and Sweet R. A new design for airway management training with mixed reality and high fidelity modeling. Medicine Meets Virtual Reality 22: NextMed/MMVR22, 220:359, 2016.

[7]. Carlson JN, Das S, De la Torre F, Callaway CW, Phrampus PE, and Hodgins J. Motion capture measures variability in laryngoscopic movement during endotracheal intubation: a preliminary report. Simulation in Healthcare: Journal of the Society for Simulation in Healthcare, 7(4):255, 2012. [PubMed: 22801254]

[8]. Carlson JN, Das S, De la Torre F, Frisch A, Guyette FX, Hodgins JK, and Yealy DM. A novel artificial intelligence system for endotracheal intubation. Prehospital Emergency Care, 20(5):667–671, 2016. [PubMed: 26986814]

[9]. Chattopadhay A, Sarkar A, Howlader P, and Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE, 2018.

[10]. Chen L, Day TW, Tang W, and John NW. Recent developments and future challenges in medical mixed reality. In 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 123–135. IEEE, 2017.

[11]. Delson N, Sloan C, McGee T, Kedarisetty S, Yim W-W, and Hastings RH. Parametrically adjustable intubation mannequin with real-time visual feedback. Simulation in Healthcare, 7(3):183–191, 2012. [PubMed: 22333883]

[12]. DiPietro R, Lea C, Malpani A, Ahmidi N, Vedula SS, Lee GI, Lee MR, and Hager GD. Recognizing surgical activities with recurrent neural networks. In International conference on medical image computing and computer-assisted intervention, pp. 551–558. Springer, 2016.

[13]. Doughty H, Damen D, and Mayol-Cuevas W. Who's better? who's best? pairwise deep ranking for skill determination. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6057–6066, 2018.

[14]. Doughty H, Mayol-Cuevas W, and Damen D. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7862–7871, 2019.

[15]. Fard MJ, Ameri S, Chinnam RB, Pandya AK, Klein MD, and Ellis RD. Machine learning approach for skill evaluation in roboticassisted surgery. In Proceedings of the World Congress on Engineering and Computer Science, vol. 1, 2016.

[16]. Fard MJ, Ameri S, Darin Ellis R, Chinnam RB, Pandya AK, and Klein MD. Automated robot-assisted surgical skill evaluation: Predictive analytics approach. The International Journal of Medical Robotics and Computer Assisted Surgery, 14(1):e1850, 2018.

[17]. Fawaz HI, Forestier G, Weber J, Idoumghar L, and Muller P-A. Evaluating surgical skills from kinematic data using convolutional neural networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 214–221. Springer, 2018.

[18]. Fawaz HI, Forestier G, Weber J, Idoumghar L, and Muller P-A. Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks. International Journal of Computer Assisted Radiology and Surgery, 14(9):1611–1617, 2019. [PubMed: 31363983]

[19]. Fedorov A, Beichel R, Kalpathy-Cramer J, Finet J, Fillion-Robin J-C, Pujol S, Bauer C, Jennings D, Fennessy F, Sonka M, et al. 3d slicer as an image computing platform for the quantitative imaging network. Magnetic resonance imaging, 30(9):1323–1341, 2012. [PubMed: 22770690]

[20]. Foglia EE, Ades A, Napolitano N, Leffelman J, Nadkarni V, and Nishisaki A. Factors associated with adverse events during tracheal intubation in the nicu. Neonatology, 108(1):23–29, 2015. [PubMed: 25967680]

[21]. Garcia J, Coste A, Tavares W, Nuno N, and Lachapelle K. Assessment of competency during orotracheal intubation in medical simulation. British Journal of Anaesthesia, 115(2):302–307, 2015. [PubMed: 26170352]

[22]. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, and Marín-Jimenez MJ. Automatic generation and detection of highly reliablé fiducial markers under occlusion. Pattern Recognition, 47(6):2280–2292, 2014.

[23]. Ghasemloonia A, Maddahi Y, Zareinia K, Lama S, Dort JC, and Sutherland GR. Surgical skill assessment using motion quality and smoothness. Journal of Surgical Education, 74(2):295–305, 2017. [PubMed: 27789192]

[24]. Hamza-Lup FG, Rolland JP, and Hughes C. A distributed augmented reality system for medical training and simulation. arXiv preprint arXiv:1811.12815, 2018.

[25]. Hatch LD, Grubb PH, Lea AS, Walsh WF, Markham MH, Whitney GM, Slaughter JC, Stark AR, and Ely EW. Endotracheal intubation in neonates: a prospective study of adverse safety events in 162 infants. The Journal of Pediatrics, 168:62–66, 2016. [PubMed: 26541424]

[26]. He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[27]. Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. In Advances in neural information processing systems, pp. 2017–2025, 2015.

[28]. Kingma DP and Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[29]. Lea C, Flynn MD, Vidal R, Reiter A, and Hager GD. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 156–165, 2017.

[30]. Lin M, Chen Q, and Yan S. Network in network. arXiv preprint arXiv:1312.4400, 2013.

[31]. Long J, Shelhamer E, and Darrell T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440, 2015.

[32]. Malpani A, Vedula SS, Chen CCG, and Hager GD. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. In International Conference on Information Processing in Computer-Assisted Interventions, pp. 138–147. Springer, 2014.

[33]. Matava C, Pankiv E, Raisbeck S, Caldeira M, and Alam F. A convolutional neural network for real time classification, identification, and labelling of vocal cord and tracheal using laryngoscopy and bronchoscopy video. Journal of Medical Systems, 44(2):1–10, 2020.

[34]. Nguyen XA, Ljuhar D, Pacilli M, Nataraja RM, and Chauhan S. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. Computer Methods and Programs in Biomedicine, 177:1–8, 2019. [PubMed: 31319938]

[35]. Qian M, Nicholson J, Tanaka D, Dias P, Wang E, and Qiu L. Augmented reality (ar) assisted laryngoscopy for endotracheal intubation training. In International Conference on Human-Computer Interaction, pp. 355–371. Springer, 2019.

[36]. Rabbi I and Ullah S. A survey on augmented reality challenges and tracking. Acta Graphica, 24(1–2):29–46, 2013.

[37]. Rahman T, Chandran S, Kluger D, Kersch J, Holmes L, Nishisaki A, and Deutsch ES. Tracking manikin tracheal intubation using motion analysis. Pediatric Emergency Care, 27(8):701–705, 2011. [PubMed: 21811199]

[38]. Sawyer T, Ades A, Ernst K, and Colby C. Simulation and the neonatal resuscitation program 7th edition curriculum. NeoReviews, 17(8):e447–e453, 2016.

[39]. Sawyer T and Gray MM. Procedural training and assessment of competency utilizing simulation. In Seminars in Perinatology, vol. 40, pp. 438–446. Elsevier, 2016. [PubMed: 27692475]

[40]. Sawyer T, Strandjord T, Johnson K, and Low D. Neonatal airway simulators, how good are they? a comparative study of physical and functional fidelity. Journal of Perinatology, 36(2):151, 2016. [PubMed: 26583944]

[41]. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, pp. 618–626, 2017.

[42]. Sigrist R, Rauter G, Riener R, and Wolf P. Augmented visual, auditory, haptic, and multimodal feedback in motor learning: a review. Psychonomic Bulletin & Review, 20(1):21–53, 2013. [PubMed: 23132605]

[43]. Wang Z and Fey AM. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. International journal of computer assisted radiology and surgery, 13(12):1959–1970, 2018. [PubMed: 30255463]

[44]. Wang Z, Yan W, and Oates T. Time series classification from scratch with deep neural networks: A strong baseline. In 2017 International Joint Conference on Neural Networks (IJCNN), pp. 1578–1585. IEEE, 2017.

[45]. Weinberg ER, Auerbach MA, and Shah NB. The use of simulation for pediatric training and assessment. Current Opinion in Pediatrics, 21(3):282–287, 2009. [PubMed: 19381090]

[46]. Woo S, Park J, Lee J-Y, and So Kweon I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19, 2018.

[47]. Wulf G, Shea C, and Lewthwaite R. Motor skill learning and performance: a review of influential factors. Medical Education, 44(1):75–84, 2010. [PubMed: 20078758]

[48]. Xiao X, Zhao S, Meng Y, Soghier L, Zhang X, and Hahn J. A physics-based virtual reality simulation framework for neonatal endotracheal intubation. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 557–565. IEEE, 2020.

[49]. Xiao X, Zhao S, Zhang X, Soghier L, and Hahn J. Automated assessment of neonatal endotracheal intubation measured by a virtual reality simulation system. In International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2020.

[50]. Zhao S, Li W, Zhang X, Xiao X, Meng Y, Philbeck J, Younes N, Alahmadi R, Soghier L, and Hahn J. Automated assessment system with cross reality for neonatal endotracheal intubation training. In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 739–740. IEEE, 2020.

[51]. Zhao S, Xiao X, Zhang X, Li W, Meng Y, Soghier L, and Hahn J. Automated assessment system for neonatal endotracheal intubation using dilated convolutional neural network. In International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2020.

[52]. Zhou B, Khosla A, Lapedriza A, Oliva A, and Torralba A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929, 2016.

[53]. Zia A, Sharma Y, Bettadapura V, Sarin EL, and Essa I. Video and accelerometer-based motion analysis for automated surgical skills assessment. International Journal of Computer Assisted Radiology and Surgery, 13(3):443–455, 2018. [PubMed: 29380122]

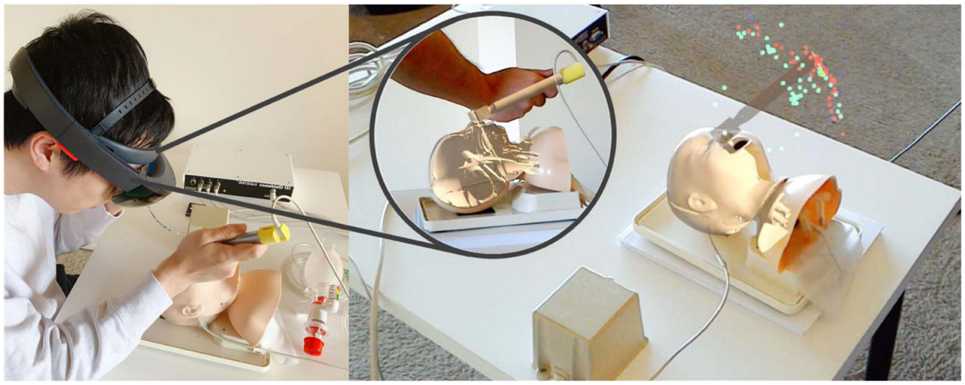


Figure 1:

Left: a trainee is conducting intubation using the system instrumented with EM sensors. Middle: the see-through visualization from the HMD provides real-time motion tracking. Right: the post-trial feedback in the playback mode shows the color-coded motion trajectory with warm colors indicating the regions that need more attention for improvement.

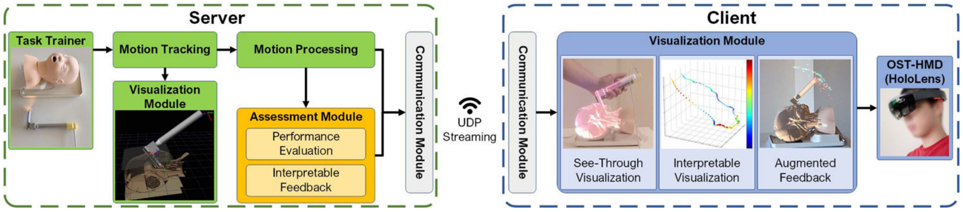


Figure 2:

The overview of our intelligent AR training framework pipeline.

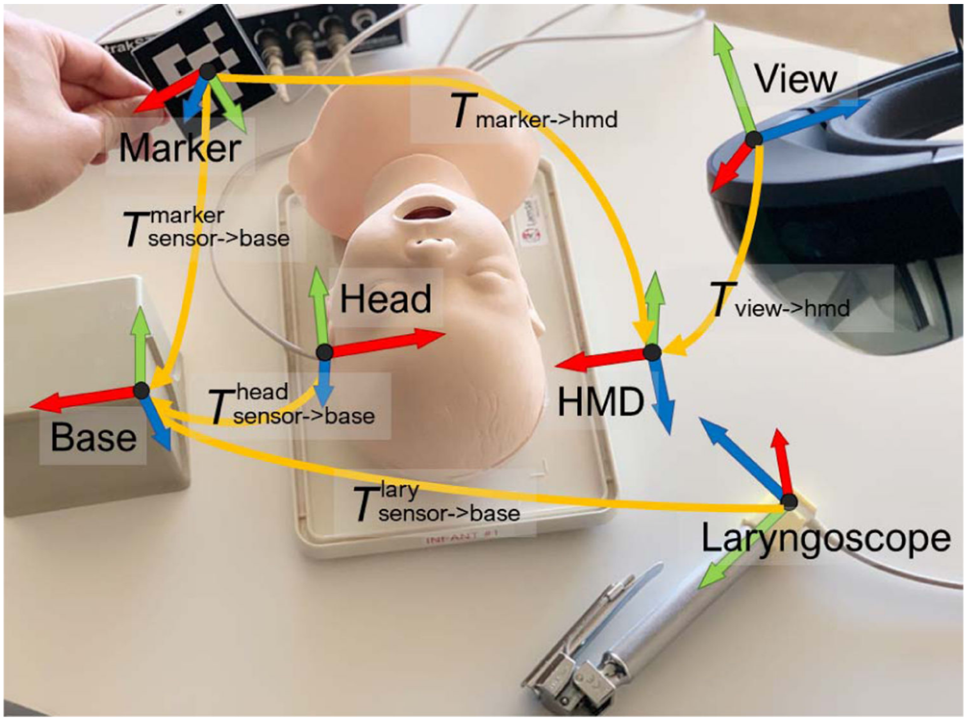


Figure 3:

The coordinates axes and the transformations between different coordinate spaces in our intelligent AR training framework.

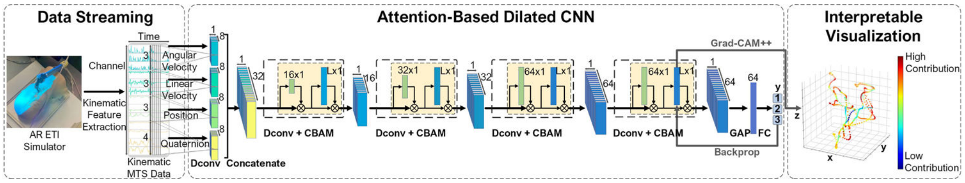
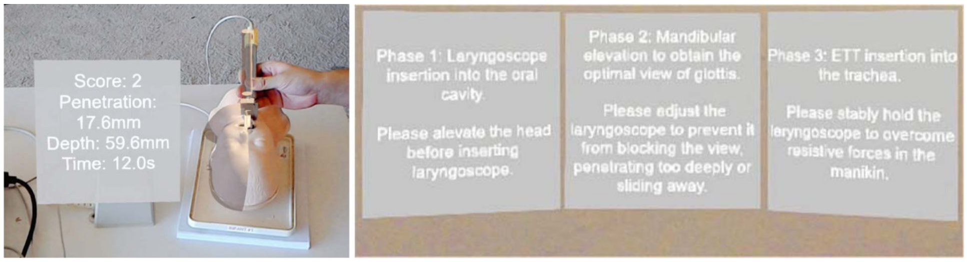


Figure 4:

The architecture of the CNN model for intelligent assessment. **DConv**, **DConv+CBAM**, **GAP**, and **FC** represent the dilated convolutional layer, the dilated convolutional block with attention module, the Global Average Pooling, and the fully connected convolutional layer, respectively. **Backprop** denotes the backpropagation to the last convolutional layer. ⊗ represents element-wise multiplication between the feature map and attention values.



Score: 2
Penetration: 17.6mm
Depth: 59.6mm
Time: 12.0s

Phase 1: Laryngoscope insertion into the oral cavity.

Please alevate the head before inserting laryngoscope.

Phase 2: Mandibular elevation to obtain the optimal view of glottis.

Please adjust the laryngoscope to prevent it from blocking the view, penetrating too deeply or sliding away.

Phase 3: ETT insertion into the trachea.

Please stably hold the laryngoscope to overcome resistive forces in the manikin.

Figure 5:

The visualization tools of our AR intelligent training framework. Left: visualization of useful features in ETI procedures. Right: visualization of instructions for each phase of the ETI procedure.

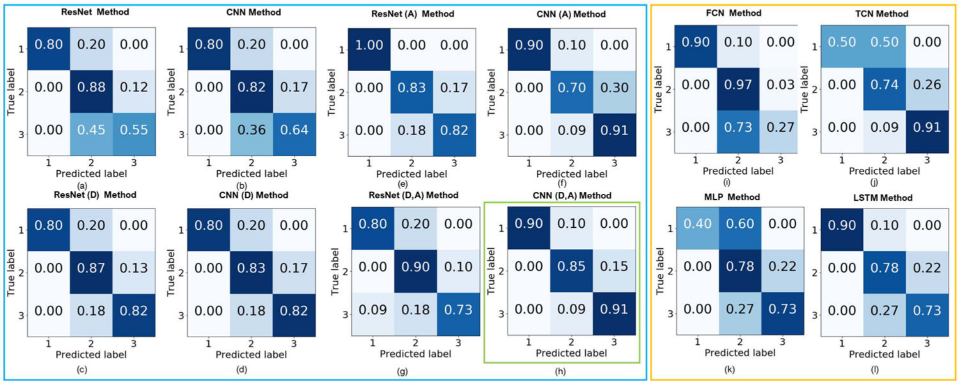


Figure 6:

The confusion matrices of different neural network models in all experiments. The matrices in the blue box are evaluated from the ablation study that includes ResNet models and CNN models, where (**A**) represents the network with attention mechanism and (**D**) represents the network with dilated convolution. The matrices in the yellow box are evaluated from the baseline comparison study. The confusion matrix of our method is shown in the green box (h).

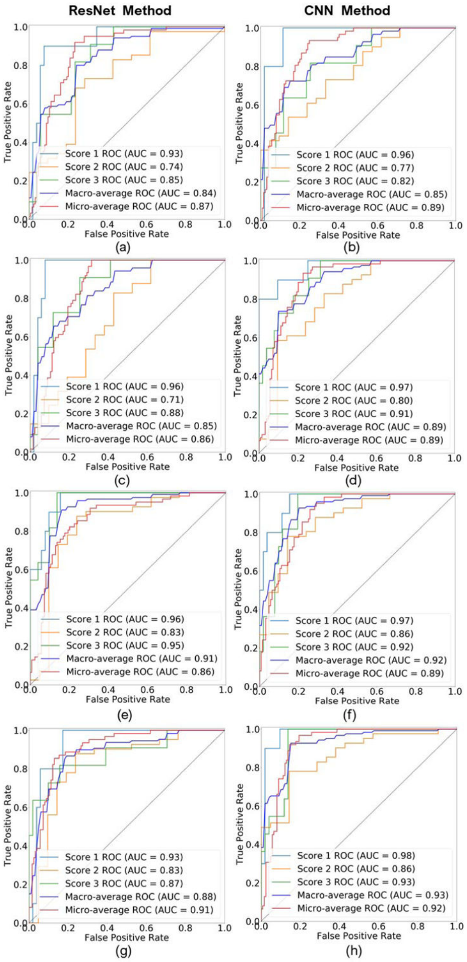


Figure 7:

The ROC curves and corresponding AUC values of different neural network models in the ablation study. The left and right columns are ResNet models and CNN models, respectively. From the top to the bottom row, the configurations are: no dilated convolution and no attention; dilated convolution but no attention; attention but no dilated convolution; and both dilated convolution and attention.

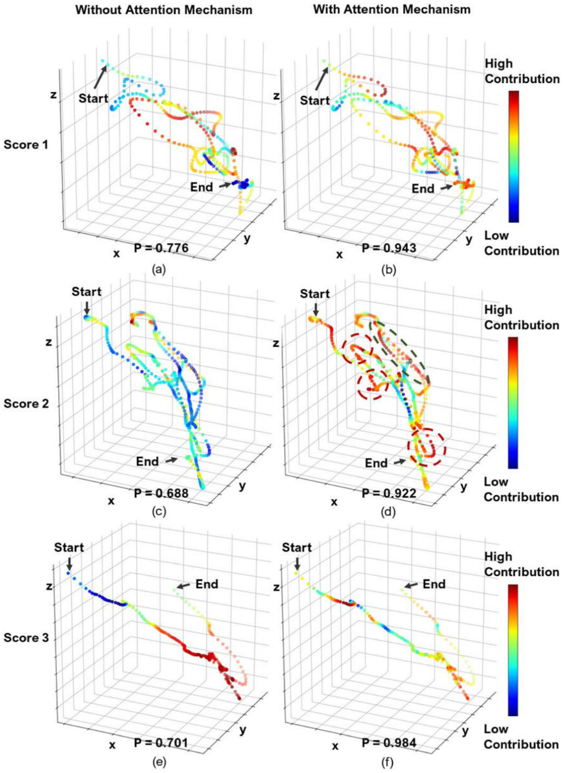


Figure 8:

Visual comparisons of all score classes with and without attention mechanism. The ground-truth score class is shown on the left of each row and P denotes the softmax score of each network for the ground-truth class. The motion regions with warmer color indicate higher contributions to the prediction.

Table 1:

A comparison of classification accuracy of various configurations. The classification accuracies are averaged over 10 repeats.

Method	**Accuracy**	
	No Dilated Convolution	**Dilated Convolution**
ResNet	79.1% (Baseline)	79.3%
CNN	78.8%	81.5%
ResNet+Attention	81.0%	83.1%
CNN+Attention	81.7%	**84.6% (Ours)**

Table 2:

A comparison of classification accuracy of different baseline methods. The classification accuracies are averaged over 10 repeats.

Author	**Method**	**Accuracy**
Fawaz et al. [17]	FCN	75.2%
Wang et al. [44]	MLP	73.9%
Lea et al. [29]	TCN	73.5%
Dipietro et al. [12]	LSTM	79.5%
Ours	Dilated CNN+Attention	**84.6%**