communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-00776-z

Deep learning for video-based assessment of endotracheal intubation skills

Check for updates

Jean-Paul Ainam[1], Erim Yanik[2], Rahul Rahul[1], Taylor Kunkes[3], Lora Cavuoto[3], Brian Clemency[4], Kaori Tanaka[4], Matthew Hackett[5], Jack Norfleet[5] & Suvranu De[2]

Abstract

Background Endotracheal intubation (ETI) is an emergency procedure performed in civilians and combat casualty care settings to establish an airway. It's crucial that healthcare personnel are proficient in these skills, which traditionally have been evaluated through direct feedback from experts. Unfortunately, this method can be inconsistent and subjective, requiring considerable time and resources.

Methods This study introduces a system for assessing ETI skills using video analysis. The system employs advanced video processing techniques, including a 2D convolutional autoencoder (AE) based on a self-supervision model, capable of recognizing complex patterns in videos. A 1D convolutional model enhanced with a cross-view attention module then uses AE features to make assessments. Data for the study was gathered in two phases, focusing first on comparisons between experts and novices, and then examining how novices perform under time constraints with outcomes labeled as either successful or unsuccessful. A separate set of data using videos from head-mounted cameras was also analyzed.

Results The system successfully distinguishes between experts and novices in initial trials and demonstrates high accuracy in further classifications, including under time pressure and using head-mounted camera footage.

Conclusions This system's ability to accurately differentiate between experts and novices instills confidence in its effectiveness and potential to improve training and certification processes for healthcare providers.

Plain language summary

Endotracheal intubation (ETI) is a medical procedure where a tube is placed into a person's windpipe (trachea) to keep their airway open. This procedure is critical in emergency and clinical settings but requires skill and experience to perform correctly. In this study, we used video analysis to assess how well ETI was performed by medical staff. Our approach involved a computer-based method that analyzed videos from multiple camera angles to evaluate ETI skills. The system could automatically distinguish between beginners and experienced professionals with high accuracy. This technology has the potential to improve medical training and certification by providing objective and automated feedback. By helping healthcare providers refine their skills, this method could lead to better patient outcomes for those undergoing ETI.

Endotracheal Intubation (ETI) is an essential airway management procedure that relies on repeated practice and timely intervention for success both in civilian and combat scenarios[1–3]. Notably, combat medics face difficulties in performing ETI, leading to failed airway management to be the second most common cause of death in the battlefield[4,5]. Thus, it is important to develop curriculums to ensure robust evaluation of healthcare providers. The gold standard in ETI skill assessment is Halstedian, i.e., an expert provides real-time or video-based post hoc feedback to the trainee[6]. This approach has several limitations, including being subjective, manual, time-consuming, and subject to poor inter-rater reliability[2,5].

Recently, video-based assessment (VBA) has received much attention in skill evaluation and education[7–9]. The advantage of VBA is that the experts can prioritize the trainee during sessions while leaving comprehensive post hoc feedback later via the video data streams. Furthermore, in recent years, deep neural networks (DNNs) have achieved substantial results in video-based tasks, addressing manual and subjective assessment, especially in related fields such as surgery[10–15]. These frameworks can learn optimal features directly from complex video data and extract high-level information for classification. However, they are challenging to interpret and fail to provide spatio-temporal feedback. Moreover, these studies utilize a constant

[1]Center for Modeling, Simulation, & Imaging in Medicine, Rensselaer Polytechnic Institute, New York, NY, USA. [2]Florida Agriculture & Mechanical University—Florida State University College of Engineering, Tallahassee, FL, 32310, USA. [3]Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, USA. [4]Department of Emergency Medicine, University at Buffalo, Buffalo, NY, USA. [5]U.S. Army Futures Command, Combat Capabilities Development Command Soldier Center STTC, Orlando, FL, 32826, USA. e-mail: sde@eng.famu.fsu.edu

camera angle, and there is a gap in the literature regarding the utility of multi-view data in skill assessment.

In the literature, multiple techniques have been proposed to analyze single-view data. These techniques can learn robust and discriminative features given video data from a single view. However, their applicability becomes limited when extended to multiple views, as they fail to learn a shared representation of the different viewpoints[16]. Existing works on Multiview data can be categorized into two groups[16]. The first group focuses on unsupervised feature extraction from multiple views using variants of auto-encoders[17–19]. They commonly employ unlabeled examples to train a multi-view DNN, then use the network as a feature extractor, followed by a standard classifier. For example, Wang et al.[19], analyzed several multi-view techniques involving an autoencoder or a paired feedforward network. They learned representations in which multiple unlabeled views of data are available at training while only one view is available for testing.

The second group of papers proposes to build a multi-view DNN for classification directly[20,21]. For instance, Ainam et al.[16] employed a multi-view DNN that exploits the complementary representation shared between views and proposed an n-pair loss function to better learn a similarity metric. In addition, Chen et al.[22] proposed solving the large discrepancy that may exist between extracted features under different views by using an asymmetric distance model. Their network also introduces a cross-view consistency regularization to model the correlation between view-specific features. Similarly, Strijbis et al.[23] proposed multi-view convolutional neural networks (MV-CNNs) for automated eye and tumor segmentation on magnetic resonance imaging (MRI) scans of retinoblastoma patients, and Kan et al.[20] proposed a multi-view task agnostic CNN that can be used directly for classification. The approach proposed by Kan et al. involves learning a view-invariant representation using a separate view-specific network for each view. However, this method becomes impractical for unbalanced and small datasets, as the limited data may lead to certain branches learning robust features while others suffer from insufficient training.

To overcome these limitations, we propose a framework that can assess clinical skills from entire videos from multiple views of the same procedure. Using cross-view information, instead of relying solely on a single view, is essential in video-based assessment of skills. With multiple camera angles or views, assessors can gain a more comprehensive understanding of the entire procedure. We exploit these differences in viewpoint and propose a pipeline that consists of a 2D autoencoder (AE) and a 1D convolutional classifier. The AE is a convolutional network built on a pre-trained self-supervised model for extracting features from entire videos of different views. The 1D convolutional network with a cross-view attention module takes such features to predict surgical skills. Inspired by the success of attention mechanisms in vision[24,25] and natural language processing[26,27], we propose a cross-view attention (xVA) that exploits the multi-view nature of the data by highlighting the most salient regions in a particular view using masks obtained from a different view. In addition, we provide visual and temporal feedback to the subjects using gradient-based class activation maps (GradCAMs)[28].

The performance of the proposed framework is tested using a dataset comprised of multi-view videos of expert and novice subjects performing ETI procedures on an airway manikin. ETI is a medical procedure for airway management to improve oxygenation in most surgeries. Studies have shown that patients who arrive at a hospital with an ineffective ETI have a lower probability of survival[29,30]. Hence, properly assessing ETI skills is fundamental to reducing complications that may increase morbidity and mortality. We tested the model's efficacy in assessing ETI skill with two classification tasks: (i) a classification analysis to separate the novice subjects from the experts and (ii) successful and unsuccessful classification of the procedure.

Hence, in this work, we explore deep neural networks for automatic and objective assessment of ETI skills using multi-view video data. We make the following contributions:

We first propose a framework with a cross-view attention module that can use the full videos from multiple views to provide objective and automated performance evaluation. Secondly, we provide a visual and temporal heatmap generated via the same DNN pipeline for informative feedback.

We propose a deep learning framework that effectively assesses endotracheal intubation (ETI) skills using multi-view video data. The system successfully differentiates between expert and novice practitioners with high accuracy and reliably classifies successful and unsuccessful intubation trials. Furthermore, we incorporate a cross-view attention module to enhance performance by leveraging multi-view information, resulting in improved classification accuracy. The model generalizes well across datasets collected in different conditions and maintains robust performance when applied to single-view data from head-mounted cameras as well as multi-view data. Additionally, the system generates visual and temporal feedback through Grad-CAM, providing interpretable insights into procedural performance. These results demonstrate the potential of automated video-based assessment in improving ETI training and evaluation.

Method

Datasets

The videos of the ETI procedure are obtained from an Institutional Review Board (IRB) and Human Research Protection Office (HRPO) approved study at the University at Buffalo where each subject was asked to perform one or more ETI procedures on an airway manikin—Life/form® Airway Larry. There are two distinct datasets that were collected.

Time-synchronized multi-view datasets from fixed cameras. These datasets were obtained using two Intel Realsense side cameras and a PTZOptics front camera at 30 frames per second. The three cameras were time-synchronized and provided fixed views of the scene. Figure 1 shows the positions of the three cameras with respect to the manikin during the procedure. The multi-view datasets comprised of two phases of the study:

Phase 1 dataset consists of three time-synchronized videos from 17 novice (5 male, 12 female) and 11 expert (7 male, 4 female) subjects. The novice subjects recruited for this study were students in healthcare-related programs with little familiarity with ETI. Experts, on the other hand, had experience ranging from one to over thirty years of practicing and teaching the ETI procedure. Each expert and novice subject performed one to five repetitions of the ETI procedure on the airway manikin. Each trial lasted a maximum of three minutes, with two minutes of rest in between trials. The dataset consists of multi-view videos from 50 successful and 24 unsuccessful trials by novices and 66 successful trials by expert subjects.

Phase 2 dataset consists of the three time-synchronized videos from 5 novice subjects (2 females and 3 males) with no overlap with the Phase 1 subjects, with a total of 31 unsuccessful and 106 successful trials. Here, placing the endotracheal tube correctly in the trachea and inflating both lungs within 3 min constitute the criteria for being labeled as Successful. While the Phase 1 dataset is used for training/testing the model, the Phase 2 dataset is used to elucidate the model's ability to generalize on an unseen dataset. Additionally, it's important to note that while data from Phase 1 were collected in two different facilities, which could have introduced bias in the classification, the data from Phase 2 do not have this issue. This explains why the Phase 2 data are considered the generalization data in this context.

Single-view dataset from a head-mounted camera. This dataset was obtained using a Tobii Pro Glasses 2 head-mounted camera at 30 frames per second from 15 novices (72 trials: 48 successful and 24 unsuccessful trials) and 8 experts (39 successful trials). This dataset is extensively more challenging than the multi-view dataset from the fixed cameras due to head motion and the lack of stability of the videos. However, head-mounted cameras can be easily deployed in simulation centers and for training combat medics without the need for complex setups involving multiple fixed cameras. Despite their challenges, head-mounted cameras provide essential aspects of the scene in front of the subjects, such as gaze behavior, finding target objects, and visual attention information. Using the head-mounted camera views tests the robustness of our model, though the cross-view attention is not used.

Figure 2 shows samples of the four different views.

All participants in the novice groups were recruited via flyers and emails distributed across the university campus, while expert providers were recruited via email through the EMS services. All participants received a $50 gift card for their participation and signed an informed consent form agreeing to participate in the study. Each participant was individually introduced to the study. To be eligible, the novices had to be at least 18 years old and have a health-related major (both undergraduate or graduate); while the experts had to be at least 18 years old and have experience performing intubations.

Network architecture

This work aims to predict the outcome of ETI skill assessment using the entire video sequences of the multiple cameras. We seek to harness the temporal and spatial information encoded in the different views to predict ETI skills. To accomplish this task, we consider two networks: a 2D denoising AE and a 1D convolutional classifier.

The encoder $\Phi : \mathbb{R}^n \rightarrow \mathbb{R}^{n_z}$ takes as input a set of video frames $\mathcal{Q} = \{q^1, q^2, \ldots, q^n\} | q^i \in \mathbb{R}^{c \times h \times w}$ and outputs a set of code vectors $\mathcal{Q}_z = \{q_z^1, q_z^2, \ldots, q_z^n\} | q_z^i \in \mathbb{R}^{n_z}$. Here, $c \times h \times w$ is the spatial dimension of the input frame and n_z is the length of the code vector. The decoder $\Psi : \mathbb{R}^{n_z} \rightarrow \mathbb{R}^n$ then takes the input frame q^i that is coded as q_z^i and outputs the reconstructed frame $\widetilde{q}^i$. We also use a pre-trained self-supervision model[31] that exposes the inner structure of the input data and constrains the decoder to output data that resembles the input data with low reconstruction loss. The pre-trained self-supervision model introduces pre-specified pixels that must be present in the output data. This also incorporates auxiliary knowledge into the model without requiring any modification of the network parameters.

The AE is trained to minimize the reconstruction loss between the input and the output and is defined as follows:

$$\mathcal{L}_{\text{DAE}} = \frac{1}{N} \sum_{i=1}^{N} ||q^i - \widetilde{q}^i||_2^2 \tag{1}$$

where N is the size of the mini-batch.

$\mathcal{L}_{\text{DAE}}$, is a mean squared error (MSE). Usually, MSE leads to blurry images and does not necessarily reflect visual similarity when comparing two images. As shown in refs. 32,33, we adopted the perceptual loss[34] and use a separate pre-trained CNN and use a distance of visual features in lower layers as a distance measure instead of $\mathcal{L}_{\text{DAE}}$ pixel-level comparison alone.

The AE is then enhanced with the perceptual loss using the features extracted from the self-supervision model and is expressed as follows:

$$\mathcal{L}_{perc}^{\phi,i}(q^i, \widetilde{q}^i) = \frac{1}{C_i H_i W_i} ||\phi(q^i) - \phi(\widetilde{q}^i)||_2^2, \tag{2}$$

where C_i, H_i, W_i are the channel, height, and width, respectively, and ϕ is the self-supervised model. The input to ϕ is a reconstructed image from the decoder and the original high-quality image. We use this loss to capture more semantic information while guiding our model to generalize. The final loss is then defined as:

$$\mathcal{L}_{final_{\text{AE}}} = \mathcal{L}_{\text{DAE}} + \mathcal{L}_{\text{perc}} \tag{3}$$

Table 1 shows the detailed architecture of the network, and Fig. 3 shows the overall framework.

The architecture of the encoder consists of eight convolutional blocks (encoder.conv2d_1 to encoder.conv2d_8). Each 2D convolution operator slides a kernel of weight over the image data and performs element-wise multiplication with the data that falls under the kernel and can be precisely described as:

$$out\left(N_i, C_{out_j}\right) = bias\left(C_{out_j}\right) + \sum_{k=0}^{C_{in}-1} weight\left(C_{out_j}, k\right) * input\left(N_i, k\right), \tag{4}$$

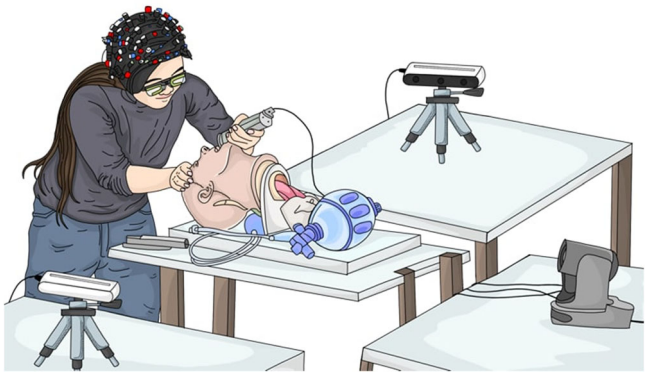

Fig. 1 | Camera positioning of the ETI task. Camera positioning during the ETI task to obtain different views of the manikin being intubated.

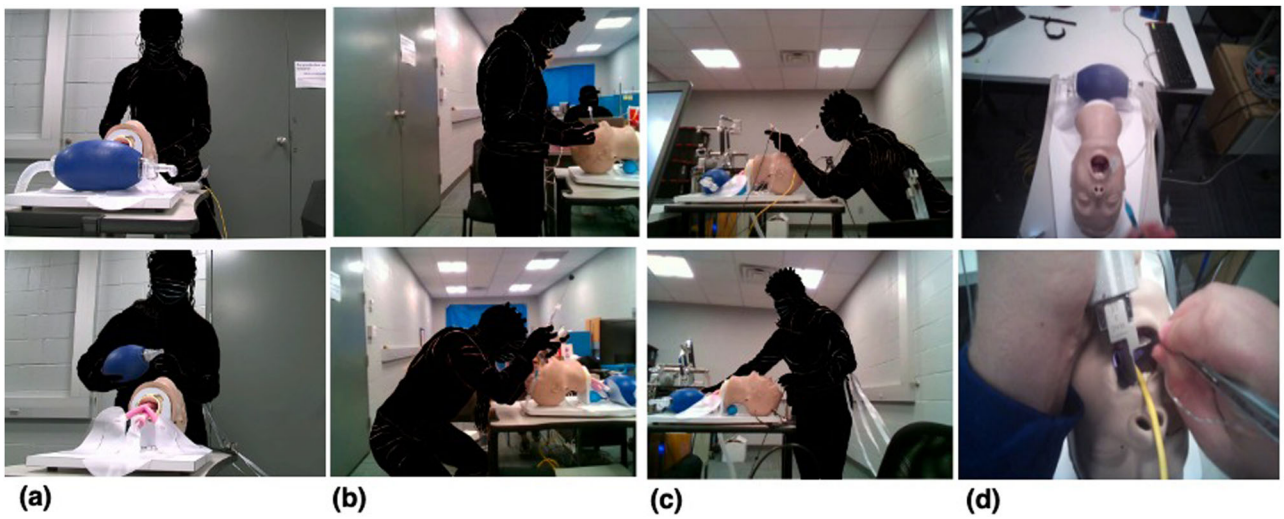

Fig. 2 | Different views for the ETI task. Four cameras are placed across the operating room and used to collect video from different angles. The first angle **a** shows the front view, the second angle **b** shows the left view, the third angle **c** shows the right, and the last angle **d** shows the head-mounted view.

Table 1 | The 2D convolutional AE network structure

Name	Input size	Kernel size	(Padding, Stride)	Output size
encoder.conv2d_1	[3 × 256 × 256]	(3 × 3)	1, 2	[32 × 128 × 128]
encoder.conv2d_2	[32 × 128 × 128]	(3 × 3)	1, 1	[32 × 128 × 128]
encoder.conv2d_3	[32 × 128 × 128]	(3 × 3)	1, 2	[64 × 64 × 64]
encoder.conv2d_4	[64 × 64 × 64]	(3 × 3)	1, 2	[64 × 32 × 32]
encoder.conv2d_5	[64 × 32 × 32]	(3 × 3)	1, 2	[128 × 16 × 16]
encoder.conv2d_6	[128 × 16 × 16]	(3 × 3)	1, 2	[128 × 8 × 8]
encoder.conv2d_7	[128 × 8 × 8]	(3 × 3)	1, 2	[128 × 4 × 4]
encoder.conv2d_8	[128 × 4 × 4]	(3 × 3)	1, 2	[128 × 2 × 2]
encoder.GAP2d	[128 × 2 × 2]	(1 × 1)	–	[128 × 1 × 1]
decoder.linear	[128]	–	–	[512]
decoder.convTrans2d_1	[128 × 2 × 2]	(3 × 3)	1, 2	[128 × 4 × 4]
decoder.convTrans2d_2	[128 × 4 × 4]	(3 × 3)	1, 2	[128 × 8 × 8]
decoder.convTrans2d_3	[128 × 8 × 8]	(3 × 3)	1, 2	[64 × 16 × 16]
decoder.convTrans2d_4	[64 × 16 × 16]	(3 × 3)	1, 2	[64 × 32 × 32]
decoder.convTrans2d_5	[64 × 32 × 32]	(3 × 3)	1, 2	[64 × 64 × 64]
decoder.convTrans2d_6	[64 × 64 × 64]	(3 × 3)	1, 2	[32 × 128 × 128]
decoder.convTrans2d_7	[32 × 128 × 128]	(3 × 3)	1, 2	[3 × 256 × 256]

After each Conv2d layer, we used a SELU activation function.

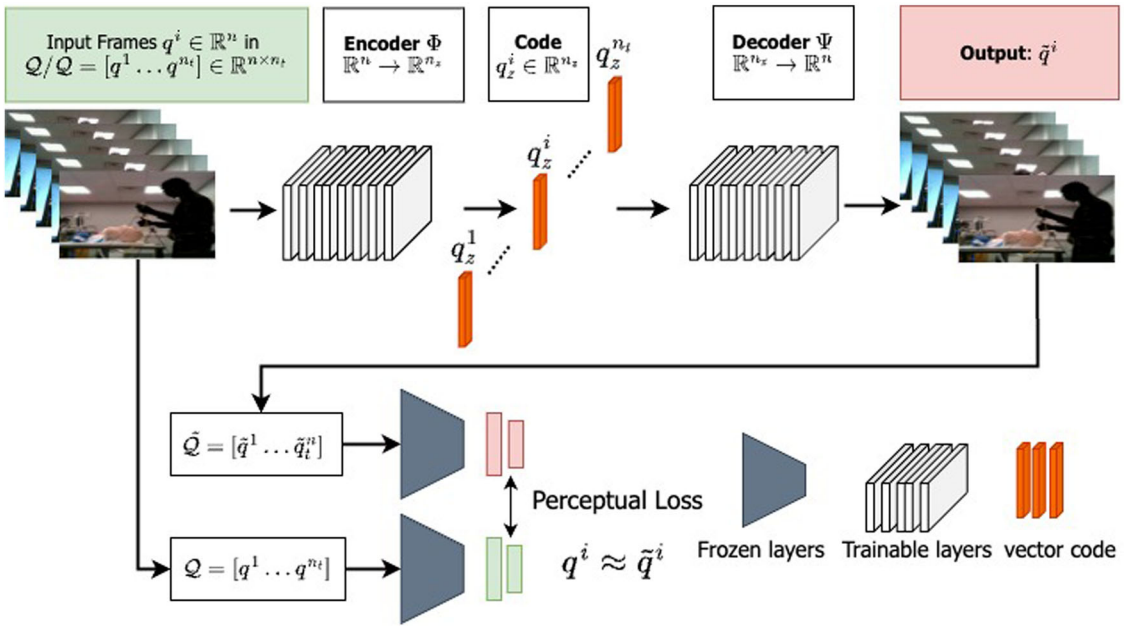

Fig. 3 | The AE framework with the self-supervised model. The AE takes as input a sequence of frames, computes a low dimensional feature representation and outputs reconstructed frames. We use the low dimensional features for classification.

where N is the batch size, C denotes the number of channels ($C = 3$ for RGB images), H is the height of input, and W is the width.

Similarly, the decoder architecture contains seven deconvolutional blocks (decoder.convTrans2d_1 to decoder.convTrans2d_7). Each block applies a 2D deconvolution operator (*a.k.a* transposed convolution operator) over the input. When initialized with the same parameters, encoder.conv2d_8 and decoder.convTrans2d_1 are inverses of each other in regard to the input and output shapes. Between the encoder and the decoder, we use a 2D global average pooling (encoder.GAP2d) to down-sample the input followed by a linear transformation (decoder.linear).

After extracting video features from the AE, we use a 1D convolutional network for classification (Fig. 4). It includes Conv1D layers, the proposed cross-view attention (*x*VA) layer, and squeeze-and-excitation layers (SE layers). The SE layers work as channel-wise attention based on Hu et al.'s[35] implementation.

The proposed cross-view attention is based on feature fusion. It is a fusion method where attention masks from one view are used to highlight the extracted features in another view. The cross-view attention is different from the self-attention mechanism. In self-attention, masks are used to highlight their features. Due to the nature of the dataset, cross-view attention that leverages multiple views is necessary. Our cross-view attention operates as follows. Given two input features $(v_1, v_2) \in R^{T \times C_{in}}$ with channels C_{in}, and temporal length T, the output $y \in R^{T \times C_{out}}$ is computed as:

$$y = \text{concat}\big([o_1, o_2]\big) \in R^{T \times C_{out}}, \tag{5}$$

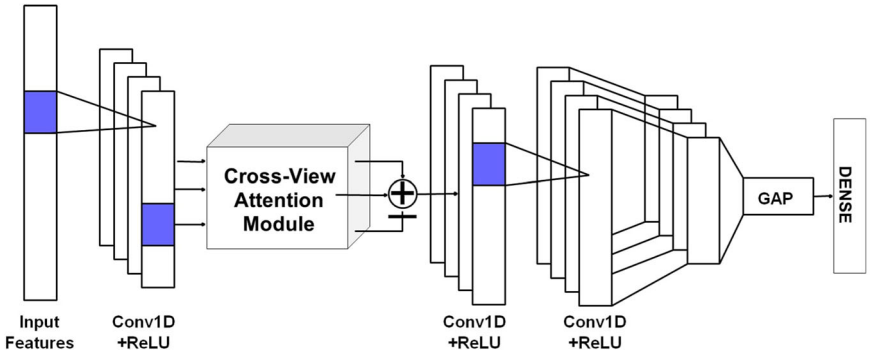

Fig. 4 | Classifier architecture. Conv1D classifier architecture. SE (Squeeze-and-Excitation network [36]) is included after each block (Conv1D + ReLU).

such that

$$o_1 = \left[m_1 \odot v_2\right] * v_1,$$

$$o_2 = \left[m_2 \odot v_1\right] * v_2,$$

$$m_1 = \sigma\left(v_1 \odot v_2\right),$$

$$m_2 = \sigma\left(v_2 \odot v_1\right),$$

where σ is the SoftMax operator, $\odot$ is the dot product, * is the matrix multiplication. Given three views $\left(v_1, v_2, v_3\right) \in R^{T \times C_{in}}$, the output o_1, o_2, o_3 corresponding to each view can be computed from the two views as follows: $v_1\left(o_1\right)$:

$$o_1 = \left[\left(m_{12} \oplus m_{13}\right) \odot v_1\right] * v_2 * v_3, \tag{6}$$

Where $m_{12} = \sigma(v_1 \odot \mathrm{v}_2)$ and $m_{13} = \sigma\left(v_1 \odot \mathrm{v}_3\right).v_2\left(o_2\right)$:

$$o_2 = \left[\left(m_{21} \oplus m_{23}\right) \odot v_2\right] * v_1 * v_3, \tag{7}$$

Where $m_{21} = \sigma(v_2 \odot \mathrm{v}_1)$ and $m_{23} = \sigma\left(v_2 \odot \mathrm{v}_3\right).v_3\left(o_3\right)$:

$$o_3 = \left[\left(m_{31} \oplus m_{32}\right) \odot v_3\right] * v_1 * v_2, \tag{8}$$

Where $m_{31} = \sigma(v_3 \odot \mathrm{v}_1)$ and $m_{32} = \sigma\left(v_3 \odot \mathrm{v}_2\right)$

The concatenated output y for the three-view can be expressed as $y = \mathrm{concat}\left(\left[o_1, o_2, o_3\right]\right) \in \mathbb{R}^{T \times 3C_{\mathrm{out}}}$. $\oplus$ represents the element-wise addition.

Generalizing to *n* views

The cross-view attention can be generalized to n views denoted as $v_1, v_2, \ldots, v_n$. Each view v_i will have an output o_i that integrates influences from all other views:

$$o_i = \left[\left(\sum_{j=1, j\neq i}^{n} m_{ij}\right) \odot \mathrm{v}_\mathrm{i}\right] * v_1 * \ldots * v_{i-1} * v_{i+1} * \ldots * v_n \tag{9}$$

Where $m_{ij} = \sigma\left(v_i \odot \mathrm{v}_\mathrm{j}\right)$ for each $j \neq i$, representing the attention mechanism between view i and view j. The concatenated output y would then be $y = concat\left(\left[o_1, o_2, \ldots, o_n\right]\right) \in \mathbb{R}^{T \times nC_{out}}$

This formulation ensures that each output o_i is influenced by all other views through a cross-view attention relative to v_i. The summation $\sum_{j=1, j\neq i}^{n} m_{ij}$ provides a way to integrate the attention weights from all other views, and the final concatenation combines the information across all views into a single feature representation. This structure allows for flexible and comprehensive interaction among multiple views.

Finally, the 1D convolutional network is trained using a cross-entropy loss:

$$L_{xent} = -\frac{1}{m}\sum_{i=1}^{m} y_i \log\left(\hat{y}_i\right), \tag{10}$$

where y_i is the ground-truth label, m is the number of classes, and $\hat{y}_i$ is the predicted label.

Baseline network. To evaluate single-view data, we present a baseline network that incorporates all components of our proposed network, excluding the cross-view attention mechanism. In other words, the baseline network takes video input from a single view and predicts the outcome. This baseline network individually assesses the left and right views and the head-mounted video.

Implementation details

The framework was implemented using PyTorch library[36]. For the AE, we used a pre-trained SimCLR[31] as the self-supervision model and fine-tuned it on our dataset. The image frames were extracted from the videos and resized to 256×256. The pixels were scaled between -1 and 1, and Gaussian noise was added to the AE input. The AE was trained for 100 epochs using Adam optimizer[37] on a batch size of 128 and with the default hyperparameters[37]. The SimCLR model parameters were frozen during the training process. To further improve the training capability, the learning rate was gradually decreased by a factor of 0.2 every 20 epochs or once learning stagnated, to a minimum value of $5e-5$.

The 1D convolutional classifier was trained on a dataset comprised of variable length sequences of a feature vector of size 32 (i.e., $n_z = 32$) representing each frame in the video. The classifier was trained for 50 epochs using the Adam optimizer on a unit batch size with default hyperparameters[37].

To compute the temporal GradCAM, we taped into the last convolution layer of the Conv1D (before the SoftMax) network to compute the gradient of the score for the given class with respect to the feature maps. Each temporal location i in the class-specific saliency map L^c is calculated as:

$$L_i^c = \sum_k w_k^c \cdot A_i^k, \tag{11}$$

where the weight $w_k^c = \frac{1}{n}\sum_i \frac{\delta Y^c}{\delta A_i^k}$, n is the sequence length and A^k is the feature map. L_i^c directly correlate with the importance of a particular temporal location (i) for a particular class c and thus functions as a temporal explanation of the class predicted by the network.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Table 2 | Results for Successful/Unsuccessful classification using the three time-synchronized fixed cameras for Phase 1 data

Successful/Unsuccessful classification		w/o *x*VA	w/o SE
Accuracy	**0.83**	0.74	0.77
Sensitivity	0.77	0.63	0.50
Specificity	0.86	0.79	0.91
F1 score	0.76	0.62	0.59
ROC AUC	0.86	0.84	0.87
MCC	0.63	0.43	0.46

Table 3 | Results for Expert/Novice classification using the three time-synchronized fixed cameras for Phase 1 data

Expert/novice classification	w/o *x*VA		w/o SE	
Metrics	**K-Fold**	**OUO***	**K-Fold**	**OUO***
Accuracy	1.0	0.98	1.0	0.98
Sensitivity	1.0	0.94	1.0	0.94
Specificity	1.0	0.90	1.0	0.93
F1 score	1.0	0.94	1.0	0.94
ROC AUC	1.0	0.93	1.0	0.93

OUO* stands for One-user-out, and K-Fold for 10-cross validation.

Table 4 | Results for Successful/Unsuccessful classification using the three time-synchronized fixed cameras for Phase 2 data

Successful/Unsuccessful classification	
Accuracy	0.92
Sensitivity	0.71
Specificity	0.98
F1 score	0.80
ROC AUC	0.93
MCC	0.76

Results

We used several metrics to evaluate the efficacity of the models, namely, accuracy, Matthews Correlation Coefficient (MCC), F1-score, sensitivity, specificity, and trustworthiness. For all evaluations, we followed a 10-fold cross-validation protocol. The data are shuffled and divided into k consecutive folds. One fold was used as the test set, while the remaining $k-1$ was the training set. Additionally, we employed a one-user-out protocol specifically for the expert vs. novice classification. This evaluation protocol was only applied to the expert vs. novice evaluation, as the data came from two different facilities. This ensures that the model learns to distinguish clinical expertise rather than individual identities.

We provide results using the baseline network, i.e., using each view of the time-synchronized multi-view datasets independently without cross-view attention in Section "Classification tasks using single camera views of the time-synchronized multi-view datasets for Phase1". In the Section "Classification tasks using all three fixed cameras for Phase 1 and 2 datasets", we introduce cross-view attention for the same datasets. Finally, in the Section "Classification tasks using the head-mounted camera," we use our model without the cross-view attention for the single-view dataset using the head-mounted camera.

Classification tasks using single camera views of the time-synchronized multi-view datasets for Phase 1

shows the classification results of the three views trained and tested on the Phase 1 dataset. We obtained an accuracy of 0.84, 0.85, and 0.72 on the successful/unsuccessful task using left, right, and front views, respectively. On the expert/novice task, we achieve much higher accuracies of 0.97, 0.94, and 0.98, respectively. These results on single views demonstrate the network's ability to learn discriminative features. However, the obtained performances remain modest compared to multi-view, as indicated in Tables 2–4. Although single views achieve interesting results, they will not provide assessors with a comprehensive evaluation. We believe that cross-view information will allow not only for improved error detection, but it will also allow assessors to identify and analyze deviations in technique by comparing the different camera angles. Consequently, to maintain accuracy classification when using multiple views, we have introduced cross-view attention as a mechanism to capture and leverage the relationships between the three views.

Classification tasks using all three fixed cameras for Phase 1 and 2 datasets

First, we present results for the Phase 1 dataset. Table 2 reports the classification results for the successful vs. unsuccessful classification.

The findings reveal that the model detects the trials with an accuracy of 0.83. The model is balanced (F1 = 0.76) in that it detects successful samples with a sensitivity performance of 0.77 while also avoiding unsuccessful ETI trials from being labeled as successful (specificity = 0.86). The receiver operator characteristics (ROC) curve and the precision-recall (PR) curve have area under the curve (AUC) values of 0.86 and 0.89, respectively, demonstrating the efficiency of our approach in distinguishing successful from unsuccessful trials.

Furthermore, we analyze the contribution of the attention modules, i.e., *x*VA and SE, to the skill classification. As seen in Table 2 the attention modules improved the accuracy via recalibrating the salient features (accuracy = 0.83/7.2% higher and F1 = 0.76/22.4% higher) and prevented overfitting to the unsuccessful samples as seen in sensitivity without the xVA and SE (18.2% and 35.1% drop), even though the separability is marginally lower (ROC AUC = 0.82/3.5% lower). This is possibly due to the overlapping distributions, which makes it harder for the classifier to differentiate the two classes, or the class imbalance, which causes the model to capture less discriminatory information for the smaller distribution. However, overall, we detect a performance improvement.

Table 3 lists performance metrics for classifying experts and novices using 10-cross validation and one-user out (OUO) protocols. The model accurately identifies all expert and novice trials in ETI tasks, with only a slight difference in performance when using the OUO protocol. Additionally, the AUC values for the ROC curve highlight the model's effectiveness in distinguishing between classes with a considerable margin. Notably, no performance loss was observed without the attention modules.

To further investigate the model's ability to generalize to new and previously unseen datasets, we use data collected during Phase 2 study. We recall that Phase 1 and Phase 2 describe the same activity with different subjects in each phase. A robust feature extraction network that successfully extracts discriminative features in Phase 1 should perform well in Phase 2, where activities are carried out in different locations. We test this, we first used the AE as a feature extractor. AE takes raw video from Phase 2 as input and output feature vectors. Then, the Conv1D classifier network is finedtuned on the extracted features to predict the trial's outcome. Here, we no longer trained AE on Phase 2 data, but use the AE model previously trained on Phase 1 data. The results of this evaluation are shown in Table 4.

Using 31 unsuccessful and 106 successful trials from Phase 2 study, we achieved a high accuracy score of 0.92 and an F1 score of 0.8, indicating perfect precision and recall. Notice that classification accuracy is better for Phase 2 data. We hypothesize that the enhanced accuracy is associated with a larger Phase 2 dataset. To substantiate this claim, we randomly sampled

50 successful and 22 unsuccessful trials from the 106 successful and 31 unsuccessful trials from Phase 2 to match those of Phase 1. We report the average results after ten runs in Table 5.

With comparable sample sizes, the model predictions are consistent between the two datasets, as illustrated in Table 5. We achieved an accuracy of 0.85 on Phase 2 data, which is comparable to the accuracy obtained in Phase 1. Moreover, all the results of the ten runs are close to the reported average, as indicated by the low standard deviation value of 0.066.

Classification tasks using the head-mounted camera

The performance results for expert/novice and successful/unsuccessful classification tasks using videos from the head-mounted camera are presented in Supplementary Table 1. The high accuracy of 0.96 and the Matthews correlation coefficient (MCC) of 0.92 highlight the ability of our model to differentiate between expert and novice trials despite the imbalance in the dataset. The network is shown to classify the unsuccessful and successful trials with an accuracy of 0.78, which is comparable to that of the classification using single static camera views, see Table 6.

Supplementary Fig. 1 shows the ROC and PR curves with the AUC measuring the degree of separability or network confidence in separating the two classes. To measure the degree of separability or network confidence in separating the two classes, we report the ROC AUC and PR AUC of the two classifications. The Expert/Novice task performs well, with an AUC of 0.99 for ROC and 0.98 for PR. Similarly, with an AUC of 0.82 for ROC and 0.67 for PR. These show the ability of our classifier to distinguish between classes. In addition, we show the trade-off between sensitivity (TPR) and specificity (1 − FPR). The results still indicate a better performance for the two tasks using head-mounted videos alone.

Discussion

In this work, we have developed a deep-learning model to assess ETI task performance. Such a model provides objective measures of the performance of clinicians and field medics, eliminating potential biases and inconsistencies in the evaluation process. However, to be acceptable to the clinical community, the model must be trustworthy, i.e., there must be confidence in its predictions. Also, a good model must be able to provide clinically relevant feedback to the trainee on their performance, without the need for manual supervision and debriefing.

We gauge the trustworthiness of our model based on metrics proposed in refs. 38,39. These metrics are developed based on the SoftMax probability of the predictions. The model is considered more reliable for a binary classification study when the SoftMax probabilities, i.e., collectively trust spectrum, are farther away from the threshold (0.5). Supplementary Fig. 2 shows the trust spectrum, i.e., the density of SoftMax probability per sample and the corresponding area under curve values, namely NetTrustScore (NTS).

The figure shows that our model has high NTS scores (>0.8) for true predictions, i.e., the spectrum is skewed towards the right, where higher SoftMax probabilities are represented for both successful and unsuccessful cases. This indicates that the model has a robust decision criterion for the true predictions, enhancing its reliability. However, the model has a trust spectrum farther away from the threshold of 0.5 for false prediction in both classes, resulting in NTSs less than 0.2. This signifies that the model can benefit from additional data on the ETI task[38].

The trustworthiness of our model for the expert vs. novice classification of ETI task via moving head-mounted camera view, yields NTS scores of 0.985 and 0.979, respectively, for the TP and TN cases (Supp. Fig. 3), which is comparable to that for the stable camera views (TP = 0.995; TN = 0.990; Supplementary Fig. 4). Notably, for the failed predictions, the NTS scores are very close to the threshold (0.5) signifying that the model does not have a strong opinion on the false predictions, a desired trait towards improving the results. However, it is important to mention that there is only one FP and FN sample. Therefore, more data are needed to solidify the model's reliability, i.e., to show that the spectrum is closer to the threshold for false predictions.

Feedback is provided using both spatial and temporal heatmaps generated using GradCAM[28]. GradCAM helps visualize parts of the ETI task that contribute the most to the classification. It generates heatmaps that highlight the regions of the input that are important for a machine learning model's prediction. Whether these highlights or explanations are equally reliable for all surgeons remains an open research question[40].

In the context of the ETI procedure, GradCAM is used to generate heatmaps that highlight the regions of a video input stream that are important for the successful placement of an endotracheal tube. By analyzing the heatmaps generated by GradCAM, clinicians can gain insights into the factors contributing to successful intubation. For example, heatmaps may show that certain features, such as head position, hand movements, and even vocal cords and tracheal rings, are particularly important for accurate tube placement. This information can help clinicians identify good posture or focus areas during the intubation procedure and may improve the overall success rate.

Additionally, these heatmaps can be used to train and refine machine-learning models designed to assist with endotracheal intubation. By using heatmaps as a form of feedback, researchers can identify the features that are most relevant, and can develop algorithms that are optimized for these features. This may lead to more accurate and reliable automated intubation systems in the future. Heatmaps generated using GradCAM are shown in

Table 5 | Successful/Unsuccessful classification results using a subset of Phase 2 data

Successful/Unsuccessful classification		
Metrics	**Phase 2**	**Phase 1**
Accuracy	0.85	0.83
Sensitivity	0.62	0.77
Specificity	0.98	0.86
F1 score	0.75	0.76
ROC AUC	0.89	0.82
MCC	0.67	0.63

(STD = 0.66).

Table 6 | Classification results using views separately

Metrics	Left view		Right view		Front view	
	Suc.vs.Uns	Exp.vs.Nov	Suc.vs.Uns	Exp.vs.Nov	Suc.vs.Uns	Exp.vs.Nov
Accuracy	0.84	0.97	0.85	0.94	0.72	0.98
Sensitivity	0.65	1.0	0.65	1.0	0.28	0.97
Specificity	0.93	0.93	0.95	0.82	0.94	1.0
F1 score	0.73	0.98	0.75	0.95	0.40	0.98
ROC AUC	0.85	0.99	0.88	0.98	0.67	0.99

Suc.vs.Uns and Exp.vs.Nov stands for successful/Unsuccessful and expert/novice tasks, respectively.

Supplementary Figs. 4–7. The heatmaps highlight the segments of the video that the network is focusing on to make a prediction.

As shown in Supplementary Fig. 4, the pose of an expert and a novice could help identify correct (or incorrect) movements that result in a successful (or unsuccessful) procedure. The expert in Supplementary Fig. 4b maintains a lowered/squatting position during the entire intubation procedure (frames 2 and 3), whereas the novice in Supplementary Fig. 4a crouches at the start (frame 2), but then stands toward the end. Proper positioning is required to obtain a good view of the airway for successful intubation. Standing prior to the end of the task may cause the provider to lose sight of the airway. The spatial heatmaps can also pick up differences in hand pose. Choking down on the laryngoscope blade (near the handle-blade connection) can lead to using the wrist as leverage, torquing the blade onto the teeth, and chipping the tooth. While the provider may successfully intubate, the patient may be harmed. The heatmap for the expert shows that the hand is placed more at the center of the handle, and the elbow is rested on the table to allow the provider to use the elbow and shoulder as the fulcrum to lift the mandible and view the vocal cords.

In Supplementary Fig. 5, we study ETI task sequences for a novice and an expert subject and visualize the temporal GradCAM and the frames associated with the activations. In this heatmap, the x-axis corresponds to the discrete sequence frames of the input video, while the y-axis represents the intensity or magnitude of the assigned weights. Hence, high values on the y-axis indicate that the frames in the input data have received higher weights, indicating their importance in the overall processing. Supplementary Fig. 5 shows that the model could detect the time sequences that contribute to the classification of the task. This feedback could explain which movements differentiate novice subjects from experts. Overall, for the novices, the weights are elevated throughout the entire procedure. The novices spend much of the task duration attempting to identify the airway and place the endotracheal tube. When the tube is perceived as in the trachea, the weights start to decrease. According to the expert, the weights are primarily elevated at the beginning of the task while identifying landmarks. Once the vocal cords are identified, the temporal heatmap weights decrease. Identifying the appropriate landmarks and vocal cords is one of the most important and challenging parts of the task. Using spatial heatmaps, we can provide constructive feedback to trainees and help them improve their performance.

We also studied the temporal heatmaps for the head-mounted cameras. For novices (see Suppl. Fig. 6), the network highlights parts of the video where the subject stands on the patient's right side and does not see the patient's head. As a result, standing sideways while performing the ETI task is an indicator used by the network to differentiate novices from experts. Moreover, unnecessary objects, such as the end of the subject's sleeve that occludes the airway manikin's face, are another indicator captured by the network. On the other hand, experts stand near the top, facing the manikin's head, and can see the airways. There is also an elevated weight from the model when the expert is checking the equipment, which is a task component stressed during training. When comparing repeat attempts for the novices, after successful intubations, the participants adjusted their body position, and the temporal heatmap weights shifted such that the elevated period was near the start and declined quickly after successful tube placement.

Similarly, Supplementary Fig. 7 shows the indicators used by the network to differentiate successful from unsuccessful trials. For successful trials, the forehead is well exposed, and the airway is visible, whereas, for unsuccessful trials, the subject is seen to apply pressure on the manikin forehead.

In this section, we used both spatial and temporal heatmaps to derive explanations from class activations. The interpretation of the heatmaps was conducted with the support of clinical team members who are experts in the performance, training, and evaluation of endotracheal intubation. However, in other cases, the conclusions may not always align with expert human explanations[40] and need to be supported by an expert explanation. GradCAM and its derived explanations are intended to complement human insights by pinpointing the elements of a video most relevant to evaluating a surgical task.

This study has several limitations. First, the modeling pipeline is not entirely automated and involves two distinct steps: a 2D autoencoder for feature extraction and a Conv1D network for classification. This configuration may restrict seamless integration and efficiency, suggesting a potential improvement by developing a fully automated and end-to-end learning model. Additionally, using a manikin that lacks realism could potentially impede participant performance. The uniform size of the manikin fails to capture the variability in human body size, consequently limiting the generalizability of our results to a more diverse population. Furthermore, the current study's findings are confined to simulated settings and do not include a thorough comparison with human raters. In future research, we will validate the model's effectiveness in actual clinical intubation scenarios involving real patients and collect a comprehensive dataset of human raters. We will aim to rigorously evaluate human raters' accuracy and consistency against our system. This comparison will ensure that our system not only meets but strives to exceed human performance standards. Each enhancement will also support our goal of providing interpretable feedback to users.

In this paper, we addressed the limitations of traditional, subjective evaluation methods using multi-view data, which captures more detailed aspects of ETI performance. It considerably improves the assessment process in critical medical procedures by distinguishing between different skill levels of practitioners and between successful and unsuccessful endotracheal intubations. The integrations of technologies like convolutional autoencoders cross-view attention modules, and GradCAMs offer precise visual feedback, enhancing training and correcting errors in ETI techniques. The potential clinical impact is substantial, particularly in environments where rapid and precise ETI is vital, as it could enhance training protocols, improve practitioner preparedness, and ultimately lead to better patient outcomes by reducing complications from poor intubation practices. This represents both a technological advancement and a major contribution to medical education and patient safety.

Data availability

All datasets used in this study, including the two-dimensional features (temporal and feature vector) and 1D feature vectors, are accessible through the authors' GitHub. Source data for the figures presented in this paper are available in the figure's subdirectory of the same repository. All other datasets are available from the corresponding author upon request.

Code availability

The computer code used for this study is also available in the authors' GitHub. The repository includes scripts for data processing, feature extraction, and model training, along with documentation for reproducing the results described in this work.

Received: 10 November 2023; Accepted: 20 February 2025;
Published online: 14 April 2025

References

1. Lim, C. et al. N. Development of a hand motion-based assessment system for endotracheal intubation training. *J. Med. Syst.* **45**, 81 (2021).
2. Zhao, S. et al. 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. 738–739. https://doi.org/10.1109/VRW50115.2020.00220 (2020).
3. Bedolla, C.N. *Airway Management: Review of Current Devices and Development of a Novel Endotracheal Tube For Emergency Combat Care*. (The University of Texas at San Antonio, 2022).
4. Zhao, S. et al. Automated assessment system for neonatal endotracheal intubation using dilated convolutional neural network. *Annu. Int. Conf. IEEE Eng. Med Biol. Soc.* **2020**, 5455–5458 (2020).
5. Xiao, X., Zhao, S., Zhang, X., Soghier, L. & Hahn, J. Automated Assessment of Neonatal Endotracheal Intubation Measured by a

Virtual Reality Simulation System. *Annu Int Conf. IEEE Eng. Med. Biol. Soc.* **2020**, 2429–2433 (2020).
6. Lim, C. et al. Multi-sensor feature integration for assessment of endotracheal intubation. *J. Med. Biol. Eng.* **40**, 648–654 (2020).
7. Pugh, C. M., Hashimoto, D. A. & Korndorffer, J. R. Jr The what? How? And who? Of video based assessment. *Am. J. Surg.* **221**, 13–18 (2021).
8. McQueen, S., McKinnon, V., VanderBeek, L., McCarthy, C. & Sonnadara, R. Video-based assessment in surgical education: a scoping review. *J. Surg. Educ.* **76**, 1645–1654 (2019).
9. Yanik, E. *Deep Learning for Video-based Assessment of Surgical Skills*. (Rensselaer Polytechnic Institute, 2022).
10. Doughty, H., Damen, D. & 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).
11. Funke, I., Mees, S. T., Weitz, J. & Speidel, S. Video-based surgical skill assessment using 3D convolutional neural networks. *Int. J. Comput. Assist. Radiol. Surg.* **14**, 1217–1225 (2019).
12. Yanik, E. et al. One-shot skill assessment in high-stakes domains with limited data via meta learning. *CBM* **174**, 108470 (2024).
13. Yanik, E., Kruger, U., Intes, X., Rahul, R. & De, S. Video-based formative and summative assessment of surgical tasks using deep learning. *Sci. Rep.* **13**, 1038 (2023).
14. Lajkó, G., Nagyné Elek, R. & Haidegger, T. Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery. *Sensors (Basel)* **21**, 5412 (2021).
15. Ming, Y. et al. Surgical skills assessment from robot assisted surgery video data, In: *2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA)*, pp. 392–396. https://doi.org/10.1109/ICPECA51329.2021.9362525 (2021).
16. Ainam, J.-P., Qin, K., Liu, G. & Luo, G. View-invariant and similarity learning for robust person re-identification. *IEEE Access.* **7**, 185486–185495 (2019).
17. Ngiam, J. et al. Multimodal deep learning, In: *Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress, USA*, pp. 689–696 (2011).
18. Srivastava, N. & Salakhutdinov, R. Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* **15**, 2949–2980 (2014).
19. Wang, W., Arora, R., Livescu, K. & Bilmes, J. On deep multi-view representation learning. In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning*. Vol. **37**, 1083–1092 (2015).
20. Kan, M., Shan, S. & Chen, X. Multi-view deep network for cross-view classification. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4847–4855. https://doi.org/10.1109/CVPR.2016.524 (2016).
21. Su, H., Maji, S., Kalogerakis, E. & Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In: *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 945–953. https://doi.org/10.1109/ICCV.2015.114 (2015).
22. Chen, Y.-C., Zheng, W.-S., Lai, J.-H. & Yuen, P. C. An asymmetric distance model for cross-view feature mapping in person reidentification. *IEEE Trans. Circuits Syst. Video Technol.* **27**, 1661–1675 (2017).
23. Strijbis, V. I. J. et al. Multi-view convolutional neural networks for automated ocular structure and tumor segmentation in retinoblastoma. *Sci. Rep.* **11**, 145–190 (2021).
24. Xu, H. & Saenko, K. Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: *Computer Vision—ECCV 2016*, pp. 451–466 (Springer International Publishing, Cham, 2016).
25. Yang, Z., He, X., Gao, J. Deng, L. & Smola, A. Stacked attention networks for image question answering. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21–29. https://doi.org/10.1109/CVPR.2016.10 (2016).
26. Lu, J., Yang, J., Batra, D. & Parikh, D. Hierarchical question-image co-attention for visual question answering. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems.* (Curran Associates Inc., USA, 2016) pp. 289–297.
27. Vaswani, A. et al. Attention Is All You Need. In: *NIPS* (2017).
28. Selvaraju, R.R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626. https://doi.org/10.1109/ICCV.2017.74 (2017).
29. Rodriguez, J.J. et al. Meta-analysis of failure of prehospital endotracheal intubation in pediatric patients. *Emerg. Med. Int.* https://doi.org/10.1155/2020/7012508 (2020).
30. Sirbaugh, P. E. et al. A prospective, population-based study of the demographics, epidemiology, management, and outcome of out-of-hospital pediatric cardiopulmonary arrest. *Ann. Emerg. Med.* **33**, 174–184 (1999).
31. Chen, T., Kornblith, S., Norouzi M. & Hinton G. A simple framework for contrastive learning of visual representations, In: *International Conference on Machine Learning*, pp. 1597–1607 (2020).
32. Ulyanov, D., Vedaldi, A. & Lempitsky, V. It takes (only) two: adversarial generator-encoder networks. *AAAI*. https://github.com/DmitryUlyanov/AGE. Accessed April 28, 2024 (2018).
33. Makhzani, A. et al. Adversarial autoencoders. *International Conference on Learning Representations* (2016).
34. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In: (eds Leibe B., Matas J., Sebe N., Welling M.) *Computer Vision—ECCV 2016, Springer International Publishing, Cham*, pp. 694–711 (2016).
35. Hu J., Shen L. & Sun G. Squeeze-and-excitation networks, In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745 (2018).
36. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library, In: *Advances in Neural Information Processing Systems 32, Curran Associates, Inc.*, pp. 8024–8035 (2019).
37. Kingma D.P., Ba J., Adam: {A} Method for Stochastic Optimization, In: 3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings. (2015).
38. Hryniowski A., Wang X.Y. & Wong A. Where does trust break down? A quantitative trust analysis of deep neural networks via trust matrix and conditional trust densities. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.14701 (2020).
39. Wong, A., Wang, X.Y. & Hryniowski A. How much can we really trust you? Towards simple, interpretable trust quantification metrics for deep neural networks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2009.05835 (2020).
40. Xue, Y. et al. An AI system for evaluating pass fail in fundamentals of laparoscopic surgery from live video in realtime with performative feedback, In: *Proceedings—2023 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Institute of Electrical and Electronics Engineers Inc.*, pp. 4167–4171. https://doi.org/10.1109/BIBM58861.2023.10385428 (2023).

Acknowledgements

We gratefully acknowledge the support of this work through the U.S. Army Futures Command, Combat Capabilities Development Command Soldier Center STTC cooperative research agreement #W912CG-21-2-0001. The authors would like to thank Dr. Lora Cavuoto and her group for data collection and the attending subjects for their dedication to this study.

Authors contributions

Conceptualization, methodology, writing original draft: J.P.A. and E.Y. Supervision, writing original draft and verification: R.R., T.K., L.C., and S.D. Supervision, validation revision and dataset collection: B.C. and K.T. Validation and supervision: M.H. and J.N.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00776-z.

Correspondence and requests for materials should be addressed to Suvranu De.

Peer review information *Communications Medicine* thanks Arunabha Karmakar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025