2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Auxiliary Tasks Based Framework for Automated Medical Skill Assessment with Limited Data

Shang Zhao[1], Xiaoke Zhang[2], Fang Jin[2] and James Hahn[3]

Abstract— **Automated medical skill assessment facilitates medical education by merging varying clinical experiences across instructors for standardizing medical training. However, medical datasets for training such automated assessment rarely have satisfactory sizes due to the cost of data collection, safety concerns and privacy restrictions. Current medical training relies on evaluation rubrics that usually include multiple auxiliary labels to support the overall evaluation from varying aspects of the procedure. In this paper, we explore machine learning algorithms to design a generalizable auxiliary task-based framework for medical skill assessment to address training automated systems with limited data. Our framework exhaustively mines valid auxiliary information in the evaluation rubric to pre-train the feature extractor before training the skill assessment classifier. Notably, a new regression-based multi-task weighting method is the key to pre-train a meaningful feature representation comprehensively, ensuring the evaluation rubric is well imitated in the final model. The overall evaluation task can be fine-tuned based on the pre-trained rubric-based feature representation. Our experimental results on two medical skill datasets show that our work can significantly improve performance, achieving 85.9% and 97.4% accuracy in the intubation dataset and surgical skill dataset, respectively.**

I. INTRODUCTION

Medical education plays an important role in providing healthcare providers with rich domain knowledge and proficient medical skills. It is critical to have supervision from instructors during the training session to improve the training quality. However, limited training opportunities are offered due to instructors' heavy clinical duties, which hinder skill acquisition. With the increasing need for healthcare providers, there is a necessity to develop a new training paradigm that facilitates learning efficiency. Disruptive technologies, such as neural networks, have been used to mitigate the conflict between the limited efficiency of training healthcare providers and the pressing demand for proficient healthcare providers.

Machine learning-based performance analysis has become a promising approach to detect discriminative patterns from the performance data to deliver statistical information-based inference for varying complex skill evaluations, such as evaluating surgical skills [9] and neonatal endotracheal intubation (ETI) skills [15]. However, the power and reliability of machine learning models are limited by the acquisition of sufficient data. Collecting more medical data is a common solution but is time-consuming and expensive. Therefore, it is necessary to develop a new framework to train effective neural networks with limited data by utilizing all aspects of information collected, such as taking advantage of the evaluation rubric in the medical skill training.

Neural networks, such as convolutional neural networks (CNN), offer implicit feature extraction and feature selection, requiring less manual intervention than traditional methods [1]. Fawaz et al. [4] developed fully convolutional networks (FCN) with kinematic multivariate time-series (MTS) data for skill evaluation. Dipietro et al. [3] trained a long short-term memory (LSTM) model for gesture recognition. However, it is difficult to deploy these models for medical training because of insufficient training data. Supervised learning with limited training data is prone to reaching an undesired convergence that has poor generalization.

Data augmentation and **representation learning** are two representative strategies to address the undesired convergence with limited data. **Data augmentation** expands the dataset with new samples by applying appropriate transformations on the original data. Wang et al. [13] partitioned every motion sequence into small segments and assigned the corresponding sequence label to those fragments. However, data augmentation only increases the sample size but cannot introduce unseen patterns in the dataset. **Representation learning** with unlabeled data saves resources on data labeling, which learns the potential new patterns implicitly. Dipietro et al. [3] pre-trained the model with motion imitation as an auxiliary task to overcome the shortage of training data. The auxiliary tasks learn pattern distributions that contribute to the primary task to produce better convergence [11]. But unsupervised auxiliary tasks cannot contribute to detecting valid patterns that matter in overall skill evaluation, causing the convergence's uncertainty. Van et al. [12] designed multi-task networks for progress prediction and gesture recognition so that the model can be further regularized by varying downstream tasks to prevent overfitting. But they did not explore the weights of tasks, which is a critical factor that affects the training quality.

Additionally, there is little research that imitates the entire evaluation rubric of medical skills to model the performance evaluation. The focus of this paper is to transfer the evaluation rubrics into training an automated framework. Evaluation rubrics have two common features. **Score-based assessment tools** include overall performance score and a set of auxiliary scores that describe the specific characteristics

*Research reported in this publication was supported by NIH R01HD091179.

[1]Shang Zhao is with the Department of Computer Science, George Washington University, USA edwinz@gwmail.gwu.edu

[2]Xiaoke Zhang and Fang Jin are with Faculty of Statistics Department, George Washington University, USA xkzhang@gwu.edu and fangjin@gwu.edu

[3]James Hahn is with Faculty of Computer Science Department, George Washington University, USA hahn@gwu.edu

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

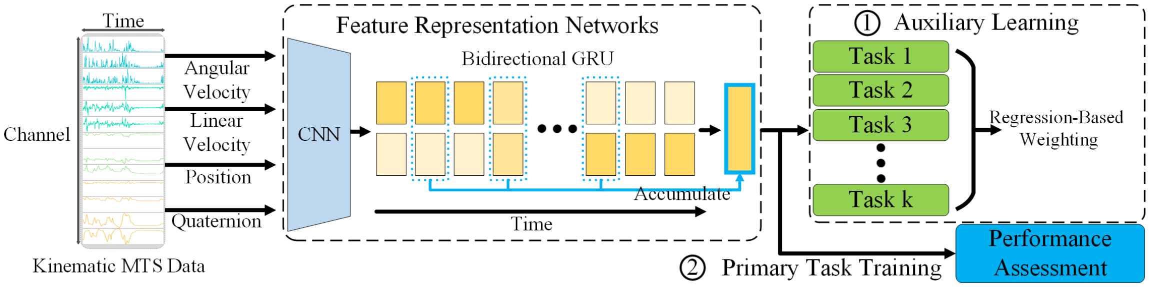

Fig. 1. The proposed auxiliary based training framework for medical skill assessment.

in the medical procedure to support the evaluation of the overall score [7], [8]. **Phase-based evaluation** evaluates the overall performance by phases, which means that any undesirable movements in key phases significantly deteriorate the outcome.

In this paper, we propose an auxiliary tasks-based automated assessment framework to imitate evaluation rubrics. The framework has the potential to be a standard way to transfer traditional medical skill evaluation to an effective automated model. The technical contribution is three-fold:

- A key phase integration-based bidirectional gated recurrent unit (bGRU) network is developed to mimic the phase-based evaluation rubric, which aggregates the contribution of summative temporal patterns of varying phases.
- An auxiliary task-based CNN-bGRU framework is proposed to learn a generalizable feature space to reduce the amount of data for training the primary task by pretraining with varying auxiliary score labels from the evaluation rubric.
- A novel regression-based weighting scheme is developed based on the correlations between auxiliary score labels and the overall score, which intentionally converges the feature space close to the primary task distribution.

II. Auxiliary Tasks Based Training Framework

A. Problem Formulation

The proposed framework includes three components: a shared feature extractor, a bundle of auxiliary task classifiers for auxiliary labels, and a primary task classifier for skill evaluation. The input of the network is kinematic MTS data of medical instruments and the output is the score categories of each task. The kinematic MTS data contain rotation (represented as quaternion), position, linear velocity and angular velocity of the medical instrument. There is an additional degree of freedom (DoF) associated with the instrument state (e.g., clipper angle for suturing) for surgical data. Given the kinematic MTS data of the instrument $\mathbf{M}_{c,t}$, where $c \in [0, C)$ and $t \in [0, T)$ denote the index of kinematic feature channels and the frame timestamp respectively. We design downstream tasks based on the corresponding auxiliary labels in the evaluation rubric of each input data, including the global overall score $\mathbf{P}_o$ and several auxiliary scores $\mathbf{A} = \{\mathbf{A}_1, ..., \mathbf{A}_N\}$, where N is the number of auxiliary tasks. Each auxiliary score $\mathbf{A}_i$ describes an important motion characteristic in the evaluation rubric, where $i \in [1, N]$. This work aims to develop a training framework to find the optimal mapping $\mathbf{f}_{\boldsymbol{\Theta}}(\cdot) : \mathbf{M} \to \mathbf{P}_o$ based on the pre-trained feature representation $\mathbf{f'}_{\boldsymbol{\Theta}}(\cdot) : \mathbf{M} \to \mathbf{A}_i$, where $\boldsymbol{\Theta}$ represents the trainable parameters.

B. Representation Learning with Auxiliary Tasks

The key factor in generating an effective automated assessment model is learning a meaningful representation space that encodes the discriminative patterns to classify varying performance levels. This requires a powerful feature extractor to generate representation features by learning from meaningful data. But small medical datasets would benefit from exploring additional valid information from auxiliary tasks besides the information from the primary task to ensure sufficient valid information. Otherwise, the model is prone to be unreliable due to the lack of training information. To this end, representation learning facilitates pattern detection from comprehensive information, resulting in a better generalizable feature representation that reduces the overfitting effects. Auxiliary tasks may contain information which correlates with the primary task labels so that auxiliary learning can contribute to the training procedure. On the other hand, auxiliary tasks with partial correlations to the primary task provide natural regularization constraints during training.

The proposed framework (Fig. 1) includes a shared feature extractor to generate a unified feature representation for all kinds of downstream tasks. Therefore, we can pre-train the shared feature extractor to learn a generalizable feature space from the valid information in varying auxiliary tasks. The shared feature extractor is developed with a CNN-bGRU encoder to learn motion patterns efficiently. CNN focuses on low-level temporal patterns of instruments and bGRU focuses on the high-level temporal patterns from different time ranges. The CNN module uses three convolutional layers to resize the low-level features to a fixed channel size. The bGRU module can detect more temporal patterns by exploring contextual information from both the past and the future. The bGRU hidden features at different key phases describe the summative patterns that end at varying timescales. Based on the phase-based evaluation in medical skills, a series of inappropriate behaviors in any phase of the procedure can ultimately lead to a serious evaluation outcome. Therefore, we aggregate summative temporal patterns at varying key phases by using uniformly selected frames. This improves the importance of summative patterns at varying phases in predicting the overall performance, imitating phase-based evaluation in medical training rubrics.

The multi-task loss function in the pre-training stage is the weighted sum of the cross-entropy loss of each auxiliary task, as shown below:

$$L(\mathbf{\Theta}, \mathbf{M}, \mathbf{A}) = \sum_{i}^{N} (\mathbf{w}_i * \mathbf{MCE}(\mathbf{f'}_{\mathbf{\Theta}}(\mathbf{M}), \mathbf{A}_i)), \quad (1)$$

where $\mathbf{MCE}(\cdot, \cdot)$ denotes the mean cross entropy loss function. After pre-training, the parameters in the pre-trained shared feature extractor $\mathbf{f'}_{\mathbf{\Theta}}$ initializes the corresponding parameters $\mathbf{f}_{\mathbf{\Theta}}$ in the final automated assessment model for training.

C. Regression-Based Weighting Scheme

The weights for evaluating total loss are critical in multi-task learning. Most existing methods only use empirical weighting schemes, such as uniform weights or manually tuned weights by parameter searching [6]. Moreover, few weighting methods have been attempted on MTS data. Empirical weighting schemes are prone to converging at a biased feature representation in the pre-training when training data distribution is different from testing data. The biased distribution leads to the negative transfer from pre-training because the representation cannot emphasize discriminative information that matters in evaluating the primary task. Therefore, it is necessary to define a mathematical correlation between the distribution between the auxiliary labels and the primary label to ensure positive transfer during training. The auxiliary labels in the medical skill evaluation rubric naturally include clear relationships with the overall performance, describing partial considerations for evaluating the overall performance. Benefiting from the partial correlation with the primary task in auxiliary tasks, additional valid information that exists in each auxiliary has the potential to learn a comprehensive representation. Deriving from the useful relationship among tasks, we develop a regression-based weighting scheme by statistically describing this relationship to guide the pre-training convergence at the closest optimal representation that facilitates the primary task training. Unlike most empirical weighting methods that lack the understanding of the training data, our method learns a feature space that converges close to the distribution of the primary task and has a generalizable representation by learning from the valid information in the auxiliary tasks.

The target of regression is to find the optimal mapping between the labels of auxiliary tasks and the label of the primary task, shown in Eq. (2). Note that the generalized estimation equation [10] is applied because of the repeated measurements for each subject in the experimental datasets and A_i terms are standardized to the same unit.

$$\mathbf{F}(\mathbf{w}, \mathbf{A}, \mathbf{P}_o) = \sum_{i}^{N} (\mathbf{w}_i * \mathbf{A}_i) \quad (2)$$

The final model's significant terms are considered as valid auxiliary tasks that have predictability for the primary task, the overall score. The coefficients of corresponding terms can be the weights for designing the total loss of the auxiliary training framework. Note that the coefficients of auxiliary labels not only model the distribution of the primary task labels from the correlated information but also use the uncorrelated information mining extra valid information that facilitates the generalization. The training will focus on the generalizable data representation to avoid overfitting on the primary task, which has better potential to reach the optimal convergence.

III. EXPERIMENTS

A. Datasets

1) Intubation Motion Dataset: The intubation dataset is derived from our previous works [17], [16], which includes a total of 478 repeated attempts from 45 subjects with varying levels of expertise. Note that a panel of expert raters scored all the intubation motions to preserve objectivity. Each motion was assigned an overall score and six feature scores: duration, motion smoothness, rocking, repositioning, force against the upper gum and insertion depth.

2) JIGSAWS Dataset: The JIGSAWS dataset [5] is a benchmark dataset for surgical skill evaluation, including 39 repeated suturing trials from 8 subjects. The score labels were rated with the objective structured assessment of technical skills (OSATS) [7], which include an overall performance score and six feature scores that evaluate skill from varying aspects, such as duration, smoothness and quality.

B. Network Architecture

All experiments used three network architectures to demonstrate the robustness of the proposed training framework and the weighting scheme, consisting of CNN [4], bGRU networks [2] and the proposed CNN+bGRU. The kernel size of CNN was 5. The hidden dimension for 2 layered bGRU was 64. All models were implemented with PyTorch 1.6 based on Python 3.7 and trained on an NVIDIA GTX 1080Ti GPU.

C. Training

Stochastic gradient descent (SGD) optimizer with cross-entropy loss was used for training. We used Leave-One-User-Out (LOUO) cross-validation on the JIGSAWS dataset. For the intubation motion dataset. We used the random partition on the intubation dataset to split 80% for training and 20% for testing because it has 6 times more subjects than JIGSAWS. The training epoch N was set to 300 and 100 for intubation and JIGSAWS datasets, respectively. The learning rate started from 1.0 and then halved every $N/6$ epochs.

IV. RESULTS

In this section, we evaluated the proposed method to demonstrate our method's effectiveness by conducting several ablation studies. There are three groups of configurations: the control group includes the models directly trained with the primary task; the reference group includes models pre-trained with the uniform weighting scheme; and the proposed regression-based weighting method.

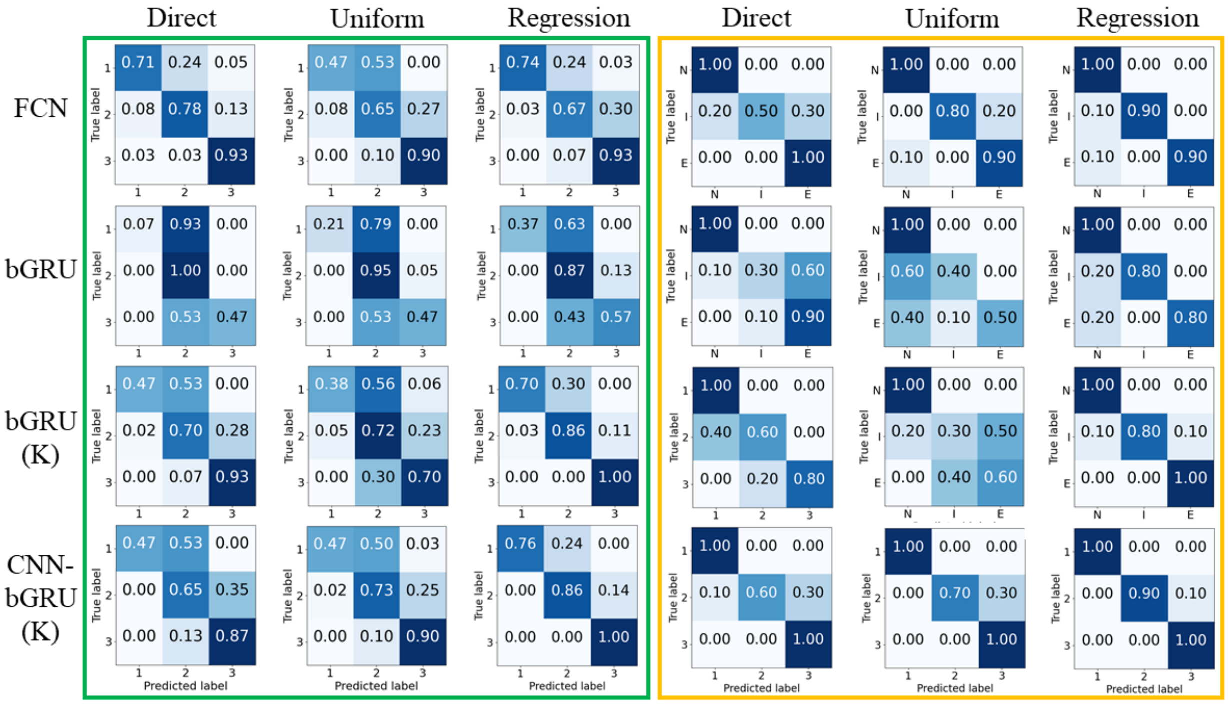

Fig. 2. The confusion matrices of intubation motion dataset (in the green box) and JIGSAWS dataset (in the yellow box) for different configurations. **K** denotes the key phase integration.

TABLE I
ACCURACY WITH DIFFERENT WEIGHTING SCHEME (INTUBATION)

	FCN	**bGRU**	**bGRU (K)**	**CNN-bGRU (K)**
Direct	81.7%	72.3%	82.3%	78.1%
Uniform	82.8%	76.3%	82.8%	82.3%
Regression	83.4%	77.8%	84.4%	**85.9%**

K=key phase integration

A. *Experimental Results on Intubation Motion Dataset*

We validated the effectiveness of the auxiliary task-based framework from the classification accuracy results (Table I). The CNN-bGRU model results from the intubation dataset show that both uniform weighting and regression weighting can achieve better performance than the control group. The result of regression weighting of CNN-bGRU is 3.6% better than uniform weights. Similar results are also observed in the results of the other two network settings. These results show that the proposed weighting scheme is able to mine more valid information than trivial methods, emphasizing the importance of utilizing supervised auxiliary labels to improve the training convergence. From the network perspective, the results of the proposed key phase integration bGRU are 7.7% better than the ones from regular bGRU on average. This shows that the key phase integration is more suitable for assessing the intubation procedure, showing some success levels on imitating the evaluation rubric. In addition, the proposed framework outperforms all the other configurations on the intubation dataset, indicating the effectiveness of the CNN-bGRU network.

To comprehensively demonstrate the proposed framework's predictability, we evaluated the confusion matrices (Fig. 2) of different configurations. The results show that the proposed regression-based weighting scheme's prediction results generally outperform others in all configurations. From the weighting scheme perspective, the results show that the regression-based weighting can evaluate the motion with more equally distributed prediction accuracy on all score classes than the other weighting configurations. One significant improvement of regression-based weighting configurations is achieving better prediction on score 1, which demonstrates the effectiveness of the proposed weighting scheme. These observations show that meaningful pre-trained representation facilitates discriminative pattern detection.

From the network perspective, regular bGRU produced biased predictions on the intubation data, which cannot distinguish the bad and good performances. In contrast, key phase integration can generate substantially better predictions on all score classes. Specifically, the prediction on score 3 is 43.0% higher than the regular bGRU with regression weighting. These results show that key phase integration can better imitate the phase-based evaluation rubric. The proposed CNN-bGRU network is substantially better in predicting score 1 than other networks with the regression-based weighting configuration. Although the proposed network did not achieve the best results, it did not produce ambiguous predictions for score 1 (bad) and score 3 (good) as CNN nor biased predictions as in regular bGRU. Furthermore, the CNN-bGRU network's performance with regression weighting is 19.0% higher in predicting score 2 than the FCN and outperforms the key phase integration bGRU on predicting score 1 by 6.0%. These results show that the proposed CNN-bGRU takes advantage of both architectures for assessing the intubation motion data.

TABLE II

ACCURACY WITH DIFFERENT WEIGHTING SCHEME (SUTURING)

	FCN	bGRU	bGRU (K)	CNN-bGRU (K)
Direct	87.2%	79.5%	84.6%	89.7%
Uniform	92.3%	71.8%	71.8%	92.3%
Regression	94.8%	89.7%	94.8%	**97.4%**

K=key phase integration

B. Experimental Results on JIGSAWS Dataset

To better demonstrate the proposed approach's generalization, we evaluated the classification accuracy and the confusion matrices for different configurations on the JIGSAWS dataset. The classification accuracy (Table II) shows that the proposed work outperforms all the other configurations by 25.6% at most. The proposed weighting scheme is substantially better than the uniform weighting. Specifically, the regression-weighting scheme in key phase integration bGRU outperforms the uniform weighting on accuracy by 23.0%. Note that the negative transfer can be observed in the bGRU configurations on the JIGSAWS dataset. But the proposed scheme always has positive effects on performance on both datasets. The effectiveness of the proposed weighting method can be observed in all networks, which further demonstrates that our method has the potential to be a generalizable solution to learn discriminative patterns from limited medical motion data.

We observed similar facts on regression-based weighting schemes from the confusion matrices (Fig. 2). Note the negative transfer with uniform weighting configurations in intermediate and expert levels. The results of uniform weighting on bGRU is 30% worse than direct training on the intermediate level, showing the importance of weighting function in supervised auxiliary learning. In contrast, the regression-based weighting scheme significantly reduces the biased predictions in all network configurations, and the CNN-bGRU improves by 30.0% on predicting intermediate level with regression-based weighting. This indicates that pre-training with our weighting scheme can stably produce a positive transfer to the primary task.

From the network perspective, key phase integration bGRU performs better in predicting experts while the regular bGRU has 20.0% false positive predictions in distinguishing expert and novice levels. This also shows that the proposed key phase integration bGRU is suitable for predicting medical skill levels. The proposed CNN-bGRU network outperforms the other networks in any training configurations, indicating the model's effectiveness. From these results, we can conclude that the proposed auxiliary task-based CNN-bGRU framework is able to assess both medical procedures with a reliable score accuracy.

V. CONCLUSION

This paper proposed a generalizable auxiliary task framework to train an automated assessment model to reduce limited medical data's side effects. The proposed key phase integration bGRU module imitated evaluation rubrics by aggregating high-level temporal patterns. A new regression-based weighting method was proposed by modeling the inter-label relationships among varying scores from the skill evaluation rubrics. Experimental results on two medical motion datasets demonstrated our method's effectiveness, which indicates that our solution has the potential to be a standard solution to automated medical skill assessment. We plan to integrate this approach to automate medical skill assessment with the mixed reality training framework [18], [14].

REFERENCES

[1] N. Ahmidi and *et al.* A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[2] J. Chung and *et al.* Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555, 2014.

[3] R. DiPietro and G. D. Hager. Automated surgical activity recognition with one labeled sequence. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 458–466. Springer, 2019.

[4] H. I. Fawaz and *et al.* 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.

[5] Y. Gao and *et al.* Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, vol. 3, p. 3, 2014.

[6] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6129–6138, 2017.

[7] J. Martin and *et al.* Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

[8] H. Niitsu and *et al.* Using the objective structured assessment of technical skills (osats) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surgery today*, 43(3):271–275, 2013.

[9] C. E. Reiley and *et al.* Review of methods for objective surgical skill evaluation. *Surgical endoscopy*, 25(2):356–366, 2011.

[10] A. Touloumis and *et al.* Gee for multinomial responses using a local odds ratios parameterization. *Biometrics*, 69(3):633–640, 2013.

[11] T. Trinh and *et al.* Learning longer-term dependencies in rnns with auxiliary losses. In *International Conference on Machine Learning*, pp. 4965–4974. PMLR, 2018.

[12] B. Van Amsterdam and *et al.* Multi-task recurrent neural network for surgical gesture recognition and progress prediction. *arXiv preprint arXiv:2003.04772*, 2020.

[13] Z. Wang and A. M. Fey. 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.

[14] X. Xiao and *et al.* 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, 2020.

[15] J. Zaichkin and G. M. Weiner. Neonatal resuscitation program (nrp) 2011: new science, new strategies. *Neonatal Network*, 30(1):5–13, 2011.

[16] S. Zhao and *et al.* Automated assessment system for neonatal endotracheal intubation using dilated convolutional neural network. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 5455–5458. IEEE, 2020.

[17] S. Zhao and *et al.* Automated assessment system with cross reality for neonatal endotracheal intubation training. In *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 739–740. IEEE, 2020.

[18] S. Zhao and *et al.* An intelligent augmented reality training framework for neonatal endotracheal intubation. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 672–681. IEEE, 2020.