International Journal of Surgery 104 (2022) 106740

Contents lists available at ScienceDirect

International Journal of Surgery

journal homepage: www.elsevier.com/locate/ijsu

Experimental Research

Intelligent cataract surgery supervision and evaluation via deep learning

Ting Wang[a,1], Jun Xia[b,1], Ruiyang Li[a,1], Ruixin Wang[a,1], Nick Stanojcic[c], Ji-Peng Olivia Li[d], Erping Long[a], Jinghui Wang[a], Xiayin Zhang[e], Jianbin Li[f], Xiaohang Wu[a], Zhenzhen Liu[a], Jingjing Chen[a], Hui Chen[a], Danyao Nie[g], Huanqi Ni[b], Ruoxi Chen[b], Wenben Chen[a], Shiyi Yin[f], Duru Lin[a], Pisong Yan[h], Zeyang Xia[i], Shengzhi Lin[j], Kai Huang[b,**], Haotian Lin[a,k,l,*]

[a] *State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Vision Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, Guangdong, China*

[b] *School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China*

[c] *Department of Ophthalmology, St. Thomas' Hospital, London, United Kingdom*

[d] *Moorfields Eye Hospital NHS Foundation Trust, London, United Kingdom*

[e] *Guangdong Eye Institute, Department of Ophthalmology, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, China*

[f] *Department of Ophthalmology, Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou, China*

[g] *Shenzhen Eye Hospital, Shenzhen Key Laboratory of Ophthalmology, Shenzhen University School of Medicine, Shenzhen, China*

[h] *Cloud Intelligent Care Technology (Guangzhou) Co., Ltd., Guangzhou, China*

[i] *Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

[j] *Guangzhou Oculotronics Medical Instrument Co., Ltd, Guangzhou, China*

[k] *Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, China*

[l] *Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China*

ARTICLE INFO

Keywords:
Cataract surgery
Artificial intelligence
Evaluation
Action recognition

ABSTRACT

Purpose: To assess the performance of a deep learning (DL) algorithm for evaluating and supervising cataract extraction using phacoemulsification with intraocular lens (IOL) implantation based on cataract surgery (CS) videos.

Materials and methods: DeepSurgery was trained using 186 standard CS videos to recognize 12 CS steps and was validated in two datasets that contained 50 and 21 CS videos, respectively. A supervision test including 50 CS videos was used to assess the DeepSurgery guidance and alert function. In addition, a real-time test containing 54 CSs was used to compare the DeepSurgery grading performance to an expert panel and residents.

Results: DeepSurgery achieved stable performance for all 12 recognition steps, including the duration between two pairs of adjacent steps in internal validation with an ACC of 95.06% and external validations with ACCs of 88.77% and 88.34%. DeepSurgery also recognized the chronology of surgical steps and alerted surgeons to order of incorrect steps. Six main steps are automatically and simultaneously quantified during the evaluation process (centesimal system). In a real-time comparative test, the DeepSurgery step recognition performance was robust (ACC of 90.30%). In addition, DeepSurgery and an expert panel achieved comparable performance when assessing the surgical steps (kappa ranged from 0.58 to 0.77).

Conclusions: DeepSurgery represents a potential approach to provide a real-time supervision and an objective surgical evaluation system for routine CS and to improve surgical outcomes.

1. Introduction

Surgical techniques vary greatly between surgeons, as do surgical outcomes [1,2]. Objective comparisons are challenging thus limiting critical evaluation and quality improvement. Moreover, surgical supervision has historically relied on senior supervision and may lack

[*] Corresponding author. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Vision Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, Guangdong, China.

[**] Corresponding author. Zhongshan Ophthalmic Center, Sun Yat-sen University, Xian Lie South Road 54#, Guangzhou, 510060 , China.

E-mail addresses: huangk36@mail.sysu.edu.cn (K. Huang), haot.lin@hotmail.com (H. Lin).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijsu.2022.106740
Received 1 March 2022; Received in revised form 16 June 2022; Accepted 16 June 2022
Available online 25 June 2022
1743-9191/© 2022 IJS Publishing Group Ltd. Published by Elsevier Ltd. All rights reserved.

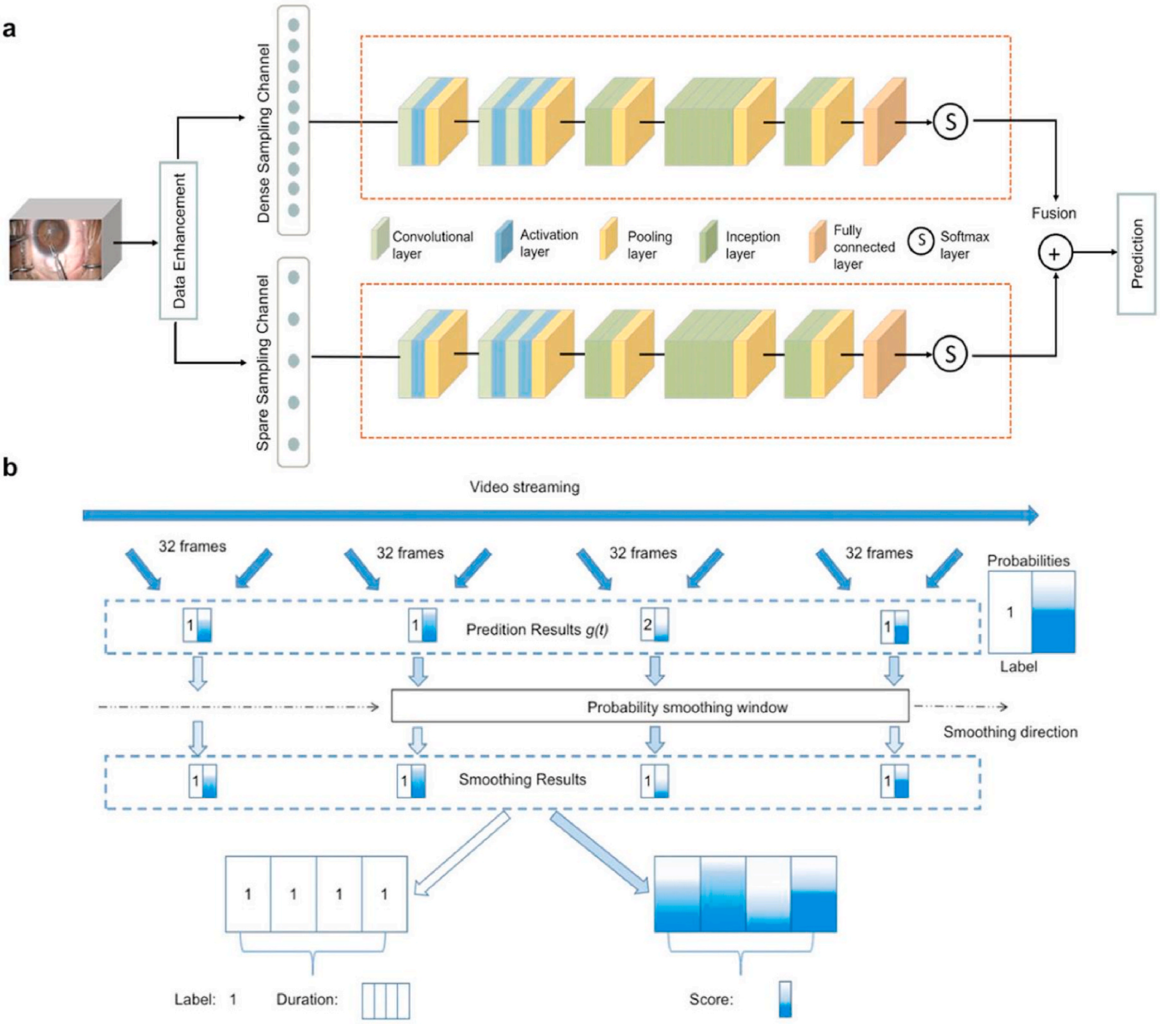

Fig. 1. A architecture of the deep convolutional neural network and assessment scheme. **a**, In our DL system development, a sampling strategy based on the fusion of sparse sampling and dense sampling is proposed. The fusion method is based on backend weighted fusion. A certain number of video frames are extracted by dense sampling and sparse sampling for training, and then the prediction error values obtained by the two methods are weighted and added. The results are taken as the final prediction error value for backpropagation. For each channel, our model architecture consisted of three convolutional layers, three activation layers, five pooling layers, nine inception layers, one fully connected layer and one softmax layer. **b**, The input data included a sequence of 32 frames in which 3D convolutional kernels operated. The probability smoothing method was used to reduce the influence of false identifications. After smoothing and rounding the result, the final step type index and its consumption times were output.

objectivity [3]. Therefore, there is an urgent need for an objective supervision and evaluation system that contributes to consistent and quality supervision globally.

Cataracts are the most common cause of vision loss and blindness in the worldwide; the World Health Organization (WHO) has estimated that the number of people with blindness globally is projected to increase from 43.3 million in 2020 to 61.0 million in 2050 [4,5]. Currently, visually significant cataracts are managed surgically [6] by employing advanced microsurgical techniques and utilizing the high-quality optics of the operating microscope. It has been reported that an annual cataract surgery (CS) rate of 4000 cases per 1 million people is needed to eliminate cataract-induced blindness [7]. Therefore, CS may be used as a representative manual surgical procedure to achieve objective assessment and standardization.

Artificial intelligence (AI) holds great promise in automated surgical phase recognition [8,9]. Automatic phase recognition is fueled by increasingly available surgical information from advanced technologies [10]. Video-based surgical records have been widely used because of their richer content [11]. The powerful algorithm of the three-dimensional convolutional neural network (3D CNN) algorithm can extract discriminant, temporal, and spatial features of a video and recognize specific actions from video streams [12–15].

In this study, we employed a 3D CNN to build an intelligent and objective system named DeepSurgery for the evaluation and supervision of surgical procedures. Because of the entire-process recognition of CS recognition process, DeepSurgery can promote CS workflow standardization during surgical training. Furthermore, with the real-time feedback of surgical supervision, DeepSurgery enables residents to accurately evaluate their surgical performance and skills, highlighting steps that need further improvement.

2. Materials and methods

2.1. Data source and preparation

Two hundred cataract surgical videos from three experienced ophthalmologists from a tertiary hospital (ZOC), who have performed over 200,000 surgeries individually. Only conventional phaco for age-related cataracts was used as an internal dataset to develop the deep learning (DL) system. Fifty CS videos from another tertiary hospital (SEH) and 21

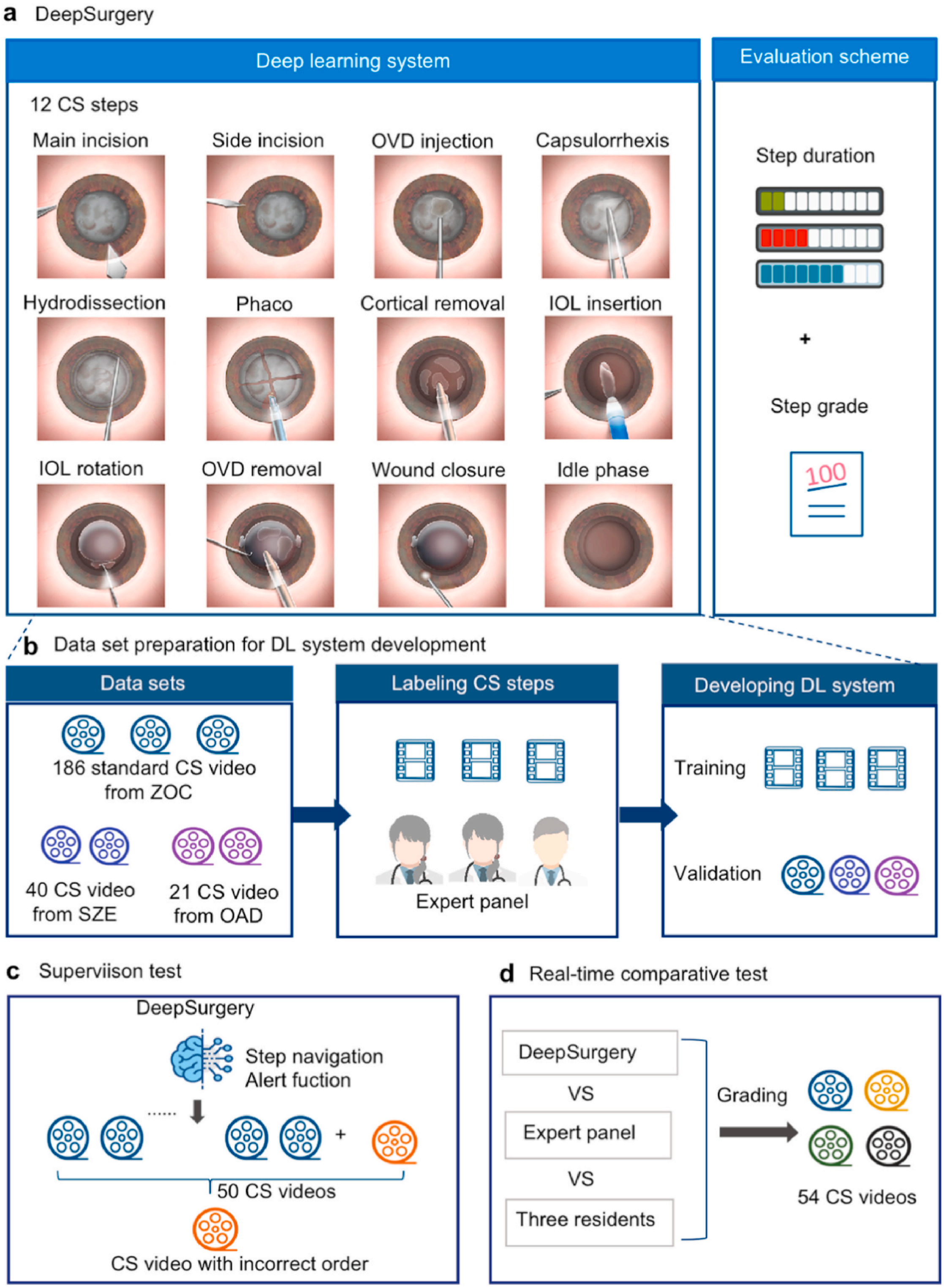

Fig. 2. Functional architecture and training pipeline of DeepSurgery. **a**, DeepSurgery includes a deep learning (DL) system and an evaluation scheme. This DL system was intended to identify 12 cataract surgery (CS) steps including the duration between two pairs of adjacent steps (idle phase). The evaluation scheme provided comprehensive step grades and step duration evaluations. **b**, The datasets included 186 standard CS videos regarding age-related cataracts from a tertiary hospital (ZOC), 50 CS videos from another tertiary hospital (SEH) and 21 CS videos downloaded from an open-access database (OAD). Each step was independently labeled by two ophthalmologists with at least 10 years of CS experience, and a senior ophthalmologist was consulted in case of disagreement (expert panel). We developed a DL system using labeled steps for training and full videos for validation. **c**, In the supervision test, DeepSurgery successfully navigated the surgical steps, identified the disordered CS video out of 50 videos, and gave timely "warning" reminders. **d**, To verify the performance of DeepSurgery in assessing surgical steps, a real-time comparative test containing 54 CSs was completed between the expert panel and DeepSurgery, and three residents. OVD: ophthalmic viscoelastic device; Phaco: phacoemulsification; IOL: intraocular lens.

CS videos downloaded from an open-access database (OAD) [16] were used for the external validation. The peak signal-to-noise ratio (PSNR) [17] was utilized to assess whether a video was blurred. If the PSNR of a video was less than 20 dB (dBs), the whole video was discarded, and thus fourteen (7%), two (4%) and zero videos were excluded from the internal dataset, the external validation from SEH and OAD, respectively). We studied 12 surgical steps [18–20]: (1) main incision formation, (2) side incision formation, (3) ophthalmic viscoelastic device (OVD) injection, (4) capsulorrhexis formation, (5) hydrodissection, (6) phaco, (7) cortical material removal, (8) intraocular lens (IOL) implantation, (9) OVD removal, (10) IOL centration and (11) wound closure through corneal hydration, and (12) idle phases.

Two ophthalmologists with at least 10 years of CS experience labeled the steps of videos and assessed step grades. Any level of disagreement was arbitrated by another senior ophthalmologist with over 15 years of CS experience. These three ophthalmologists formed an expert panel. For step grading, there were four levels: novice, beginner, advanced beginner and competent according to the International Council of Ophthalmology's Ophthalmology Surgical Competency Assessment Rubric (ICO-OSCAR: phaco) [18]. To obtain the step starting and ending times for each step (to calculate the step duration), the average values of annotations from the expert panel for each step were defined as the ground truth values.

2.2. DeepSurgery development

A total of 186 standard videos met the criteria for inclusion and were randomly divided into two parts: (1) training: 80% of the data were used to optimize the network weights and (2) tuning: 20% of the data were used to optimize hyperparameters [21].

We used a mixed enhancement method named mixup [22] for data expansion to reduce overfitting. A sampling strategy based on the fusion of sparse sampling and dense sampling was proposed for the development of our DL system. A certain number of video frames were extracted by dense sampling and sparse sampling for training, and then the prediction error values obtained by the two methods were weighted and added. The results were taken as the final prediction error value for backpropagation (Fig. 1a). The input data included a sequence of 32 frames in which 3D convolutional kernels operated. Each frame was recognized, respectively and extracted continuous features among these 32 frames. The recognition results were calculated by the neural network with a softmax function and represented as *g(t)*, where *t* is the video duration. When *g(t)* was contiguous during the test video, each step could be easily segmented. However, false identification could occur when *g(t)* was not strictly contiguous because the recognition accuracy (ACC) was not 100%. This in turn could lead to blurring of the video segmentation boundaries and inaccurate index statistics of each step. Thus, a probability smoothing window function (shown in Eq. (1)) was proposed for smoothing, where *T,* $\rho(i)$and $\rho(j)$ are the window length and given probabilities, respectively.

$$S(i) = \frac{\rho(i)}{\sum_{j=i-\frac{T}{2}}^{i+\frac{T}{2}} \rho(j)} \tag{1}$$

In most cases, false identification was associated with a low probability. The probability smoothing method reduced the weight of this identification number by a lower *ρ(i)* compared with neighboring *ρ(j)*s. In other cases, such as false identification with a probability *ρ(i)*, which was as high as *ρ(j)*, the probability smoothing method worked similar to mean smoothing (*S(i)≈ 1/T*). After smoothing *g(t)* and rounding the result, the final step type index and its consumption times were denoted as $g'(t)$ and $(t_{stepending} - t_{stepbeginning})$, $g'(t) = \sum_{\tau=1}^{T} S(\tau)g(t - \tau)$, where *T* and *t* are the window length and the video time, respectively.

The highest probability reported by the softmax function was the score awarded to the tester. Since the neural network was trained on relatively standardized CS videos, the softmax function measured the relative probabilities between different categories, and each result was considered the degree of similarity between the tester and the standard at each step. Our grading depended on the overall performance at each step, and the detailed architecture of assessment scheme is shown in Fig. 1b.

2.3. Supervision test

The CS chronological order was predetermined according to the ICO-OSCAR: phaco and briefly described as (1) main incision formation, (2) side incision formation, (3) OVD injection, (4) capsulorrhexis formation, (5) hydrodissection, (6) phacoemulsification, (7) cortical material removal, (8) OVD injection, (9) IOL implantation, (10) OVD removal, (11) IOL centration and (12) wound closure. A total of 50 CS videos (collected from the ZOC) with predetermined orders were pooled. One CS video was randomly selected from these 50 videos, the OVD injection and hydrodissection steps were deleted, and other steps were fitted together to disrupt the CS order. The dataset contained 49 CSs with predetermined order and one video with man-made disorder to test the DeepSurgery navigation and warning functions.

2.4. Real-time comparative test

Six ophthalmologists with different levels of surgical experience (Supplementary Fig. 1a) were recruited to complete 54 CSs at ZOC between January 1st, 2019, and June 30th, 2020. While the surgeries were being performed, DeepSurgery provided the names of the identified steps and the scores of the six vital steps in real time. Simultaneously, three ophthalmic residents were asked to independently complete the same test as DeepSurgery without prior information. The expert panel gave the final grading results for the six steps from 54 surgeries after consultation.

After the real-time comparative test, the three residents were asked to review these 54 CS videos and the evaluation results produced by DeepSurgery. After reviewing, they were asked to independently complete a new test: assessing another 54 CSs.

2.5. Statistical analysis

The analysis code was based on Python 3.5 under the PyTorch framework (version 0.4.0). A V100 GPU (64 GB of GPU of memory in total) and 512 GB of system memory were used in the experiments. The ACC, loss, recall, precision, and F1_score metrics were used to measure the performance of the DL system [23,24].

The intraclass correlation coefficient (ICC) was calculated to evaluate the agreement of step duration produced by DeepSurgery and the expert panel [25]. After transforming scores estimated by DeepSurgery into four grades, Cohen's kappa coefficient was calculated to assess the agreement between the grades given by the expert panel and DeepSurgery or the resident. The kappa was interpreted as follows: poor ($k < 0.00$), slight ($k = 0.00$–0.20), fair ($k = 0.21$–0.40), moderate ($k = 0.41$–0.60), substantial ($k = 0.61$–0.80) or almost perfect ($k = 0.81$–1.00) [26]. Kendall's W coefficient of concordance was calculated to assess the agreement among the three residents [27]. The Wilcoxon signed-rank test was used to compare differences in the grading results. All statistical tests were two-sided, and a *p* value less than 0.05 was considered statistically significant (R foundation for statistical computing, version 4. 0. 0).

3. Results

DeepSurgery includes a DL system and an evaluation scheme that recognizes the 12 routine CS steps, including the duration between two pairs of adjacent steps (idle phases), calculates the step duration, and grades the performance (Fig. 2a). The DL system was trained by labeled

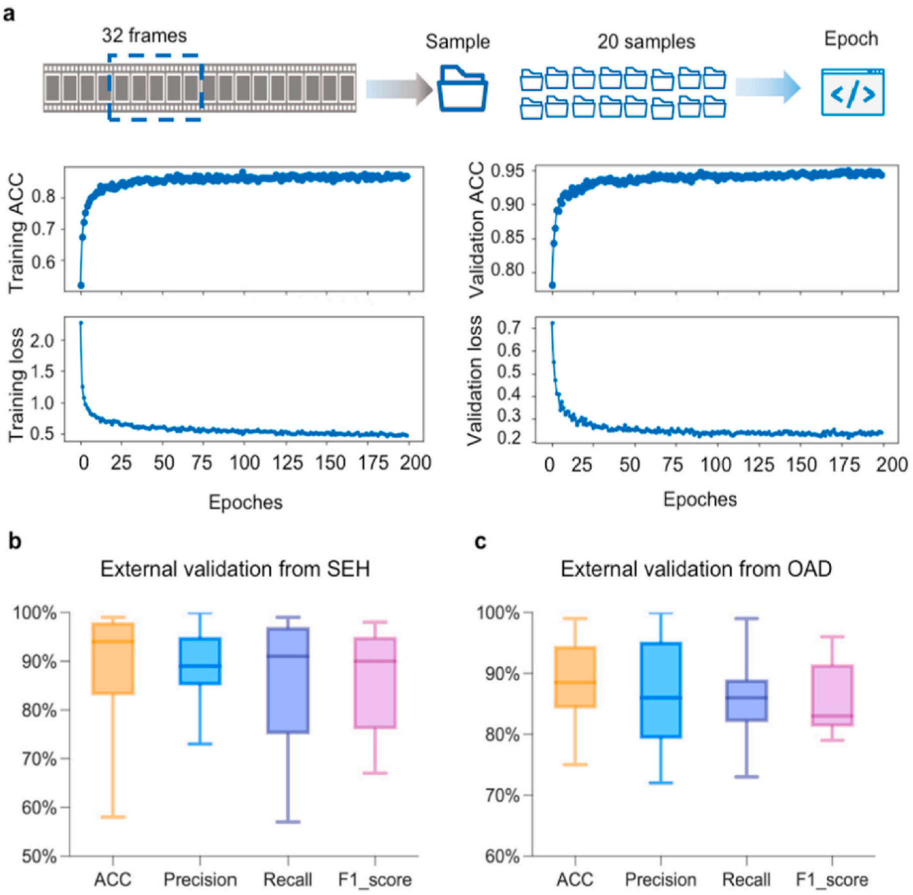

Fig. 3. DeepSurgery system performance. **a**, We defined 32 frames of the same step as one sample, and each set of 20 samples contained one epoch of training data. The ACCs on the training and validation were 85.8% and 95.06%, respectively. **b**, The ACCs of the 12 CS steps ranged from 74.70% to 98.83%, and the average ACC was 90.23% in the external validation from SEH (n = 50). The average precision, recall, and F1_score were 87.63%, 90.72%, and 91.08%, respectively. **c**, The ACCs of the 11 CS steps ranged from 58.33% to 99.40%, and the average ACC was 88.34% in the external validation from OAD (n = 21). The average precision, recall, and F1_score were 89.13%, 85.77%, and 86.74%, respectively. ACC: accuracy; SEH: a tertiary hospital; OAD: open-access database. The band shows the median, and the box indicates the middle 50% of individuals. The upper and lower error bars show the 95th and 5th percentiles of individuals, respectively.

steps and validated by full videos (Fig. 2b). In the supervision test, DeepSurgery successfully navigated the surgical step, identified the disorder CS video from 50 videos, and gave a "warning" reminder in a timely manner (Fig. 2c). To verify the DeepSurgery performance in assessing the surgical steps, a real-time comparative test containing 54 CS was completed between the expert panel and DeepSurgery and three residents (Fig. 2d).

3.1. Performance of DeepSurgery

The ACCs of the proposed system on the training and internal validation were 85.8% and 95.06%, respectively. The training process shown in Fig. 3a indicated that the neural network achieved relatively stable training results (the ACC reached 1, and the loss reached 0 smoothly). The validation ACC and loss curves showed that there was no overfitting.

The DeepSurgery performance in external validations is presented in Fig. 3b and c. Our DeepSurgery achieved stable performance for CS step detection in OAD (ACC ranged from 58.33% to 99.40%, with an average ACC of 88.34%, not including the idle phase) and SEH (the ACC ranged from 74.70% to 98.83%, with an average ACC of 88.77%). The average precision, recall, and F1_score were 89.13%, 85.77%, and 86.74% in the OAD dataset and 86.31%, 85.92%, and 85.80% in the SEH dataset, respectively (Supplementary Tables 1 and 2).

3.2. Supervision test

To further validate the ability of DeepSurgery to supervise CS and alert a surgeon to incorrect surgical steps, 50 CS videos were collected including 49 videos with a predetermined order and one edited video with incorrect orders. DeepSurgery was able to recognize each step and accurately indicate the subsequent step. DeepSurgery successfully identified the video with the incorrect orders of steps from numerous correct videos. When the OVD injection and hydrodissection steps were omitted, the surgeon was alerted (Fig. 4a and Supplementary video).

Supplementary video related to this article can be found at doi:10.1016/j.ijsu.2022.106740

3.3. Step recognition and duration measurement in the real-time test

Fig. 4b shows the stable performance for surgical step recognition in the real-time test (Supplementary Table 3). The ACCs of the 12 CS steps, including the idle phases ranged from 85.60% (wound closure) to 96.92% (phaco), and the average ACC was 90.30%. The average precision, recall, and F1_score were 92.60%, 88.24% and 89.95%, respectively.

The greatest differences between the duration measurements made by DeepSurgery and the expert panel for the phaco and cortical removal steps were 2.99 s and 2.94 s, respectively. Except for the OVD injection step (ICC = 0.55), the ICCs for the other 10 surgical steps and the idle phase were greater than 0.75 (substantial agreement). These results implied that DeepSurgery performed comparably to the expert panel in terms of measuring the durations of the surgical steps and the idle phase (Fig. 4c).

3.4. Real-time comparative test of the standard level grades

The grades of the surgeons performing these steps were provided by the expert panel according to the explanation of ICO-OSCAR: phaco (the distribution of grading is shown in Supplementary Fig. 1b) [18]. Except for that of the capsulorrhexis step, the kappa values for the other five surgical steps were greater than 0.60, which showed that our evaluation

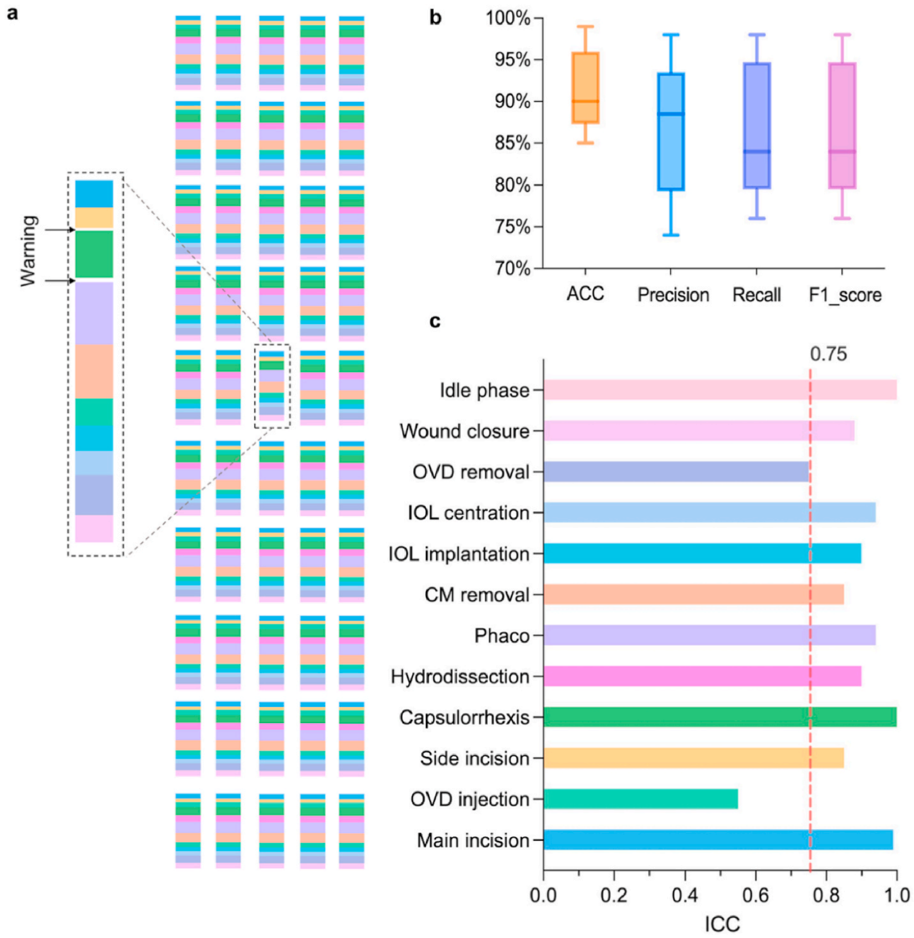

Fig. 4. Performance of the supervision test, step recognition and duration measurement of DeepSurgery in the real-time test. **a,** Fifty CS videos were collected, including 49 videos with a predetermined order and one video edited with incorrect orders. DeepSurgery successfully identified the incorrect CS video, recognized the surgical steps performed by the surgeon, and alerted the surgeon to the next step. When the OVD injection or hydrodissection steps were omitted, an alert was given. **b**, We designed a real-time test to evaluate the DeepSurgery performance. The ACCs of the 12 CS steps ranged from 85.60% to 96.92%, and the average ACC was 90.30%. The average precision, recall, and F1_score were 92.60%, 88.24% and 89.95%, respectively (n = 54). **c,** The ICC was calculated to evaluate the agreement between the step duration given by the DeepSurgery and the expert panel. Except for the OVD injection step (0.55), the ICCs of the other 10 surgical steps and the idle phase were greater than 0.75 (substantial agreement). ACC: accuracy. The band shows the median, and the box indicates the middle 50% of individuals. The upper and lower error bars show the 95th and 5th percentiles of individuals, respectively. CM: cortical material; OVD: ophthalmic viscoelastic device; Phaco: phacoemulsification; IOL: intraocular lens; idle phase: the duration between two pairs of adjacent steps.

system performed similarly to the expert panel in terms of grading these phases (Fig. 5a). The grading difference was calculated by subtracting the grades given by the expert panel from those indicated by DeepSurgery. The results showed that the majority of differences between DeepSurgery and the expert panel were one level (Fig. 5b).

To compare the CS step evaluations provided by the residents and DeepSurgery in real time, three ophthalmologic residents were asked to independently assess the 54 CS videos (Fig. 6a). Except for those of the phaco step evaluations provided by residents No. 1 and No. 3, the values for the other steps assessed by the three residents were lower than 0.6 (ranging from 0.11 to 0.54), which suggests that the residents had difficulty accurately assessing the grades of surgeons performing CS steps. In addition, substantial inconsistencies among the three residents were observed (Kendall's W < 0.6), except for the phaco step (Fig. 6b), which indicated that the three residents had different understandings of these six CS steps.

After reviewing 54 CS videos and the evaluation results given from DeepSurgery, the three residents were asked to independently complete a grading test paper containing 54 new CSs. The performance of the three residents was comparable to that of the expert panel in assessing phaco and IOL insertion steps. The agreement for a total of six vital steps between residents and the expert panel was significantly improved before and after reviewing the grading results of 54 CSs provided by DeepSurgery (Fig. 6c and d).

4. Discussion

This study established a DL system (DeepSurgery) for the evaluation and supervision of CS. DeepSurgery successfully supervised CS by providing the correct chronological order and timely alerts about incorrect workflows. In a real-time comparative test, DeepSurgery and the expert panel produced comparable results when assessing the surgical steps. To the best of our knowledge, this is the first intelligent evaluation and supervision system for CS.

Depending on the type of cataract, the required CS technique may vary greatly. With regard to standard phaco, surgical steps remain largely consistent, notwithstanding nuanced approaches depending on the maturity of the cataract and ocular and systemic comorbidities. We, therefore, used these standard steps for our research. We used 186 standard regular CS videos to train a DL system, and in the real-time comparative test, DeepSurgery exhibited robust performance. Compared to static images, high-resolution videos contain abundant and continuous information, including temporal and spatial features. In addition to the 12 surgical steps and idle phases, DeepSurgery automatically recognized the whole procedure of regular CS.

In recent years, some researchers have applied AI to CS. To identify surgical steps, many previous studies [28–30] used pictures taken from surgical videos to train models. In addition, a few studies have used surgical videos as training data. A cross-sectional study from Johns Hopkins University collected 100 CS videos to identify ten steps within the videos by using five algorithms such as CNN and recurrent neural network [31]. The unweighted ACCs corresponding to the surgical steps ranged from 0.915 to 0.959, and the areas under the receiver operating characteristic curve (AUCs) ranged from 0.712 to 0.773. The authors evaluated the ability of algorithms to perform step identification using presegmented videos, but real-life applications require algorithms that detect both segment boundaries and surgical steps. Quellec and colleagues [32] gathered 186 surgical videos; the average AUC for the joint segmentation and recognition of surgical tasks was 0.856. The motion contents of short video subsequences were modeled using

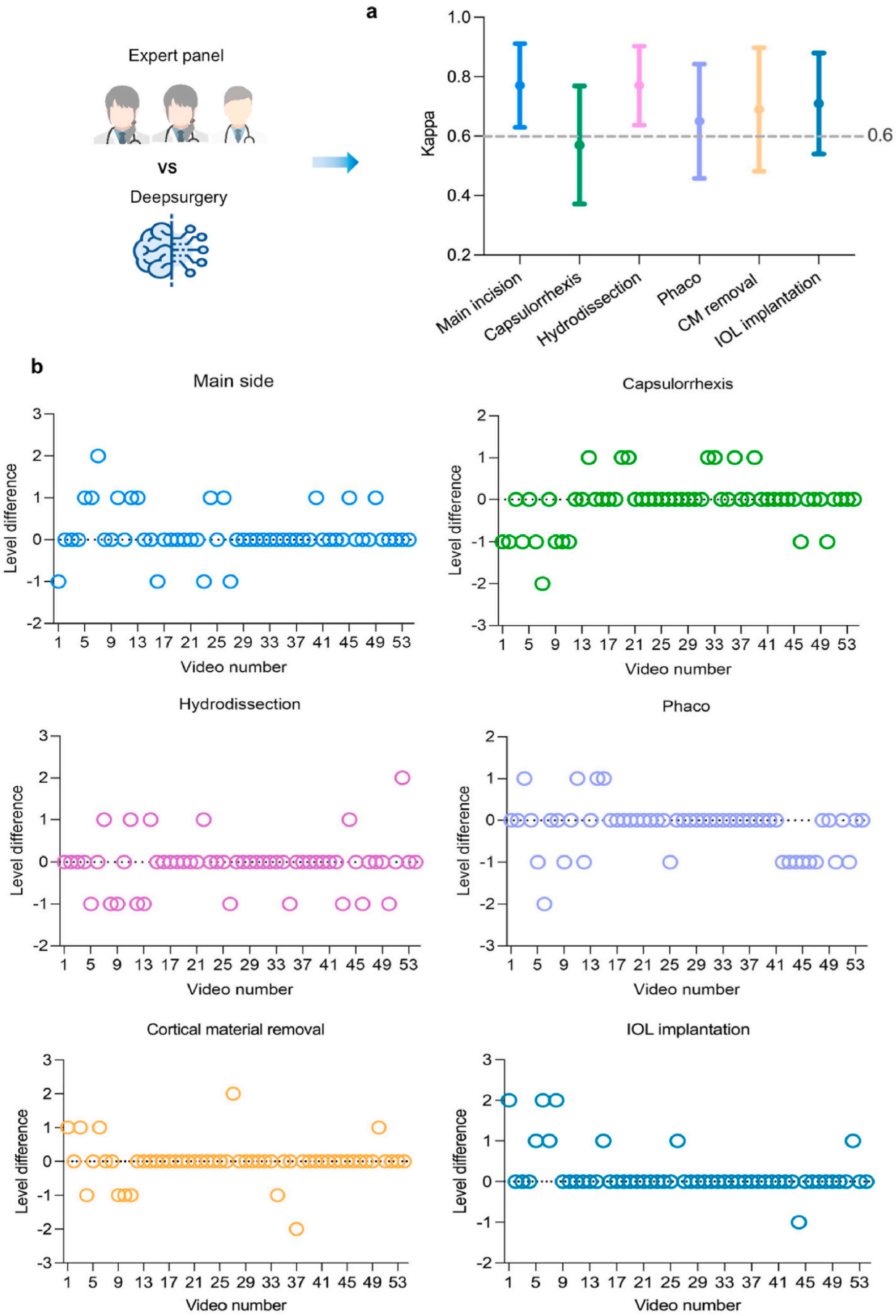

Fig. 5. The grading performance of DeepSurgery in the real-time test. **a**, Six vital surgical steps were evaluated using DeepSurgery in real time, and the grades of surgeons performing these steps were provided by the expert panel according to the explanation of ICO-OSCAR: phaco. Except for that of the capsulorrhexis step, the kappa values for the other five surgical steps exceeded 0.60, which means that the agreement between the expert panel and DeepSurgery was substantial. **b**, Detailed grading differences between the expert panel and DeepSurgery. On the y-axis, a value of zero denotes that DeepSurgery gave the same grade as that of the expert panel. A positive value indicates that the grade given by the expert panel was higher than that indicated by DeepSurgery, and a negative value denotes the opposite situation. A larger value indicates a greater difference between the results of the expert panel and DeepSurgery. The bars show 95% confidence intervals (CIs). CM: cortical material; Phaco: phacoemulsification; IOL: intraocular lens.

spatiotemporal polynomials, but the dimensions included only horizontal and vertical motion. The extraction of incomplete information from surgical procedures greatly affects the intelligent evaluation of CS steps.

Compared with previous studies, our study has several novelties. First, to the best of our knowledge, this is the first study to apply a 3D CNN for research on CS procedures that has been reported. This algorithm can extract not only the temporal but also the spatial features of videos. The sampling strategy we proposed considered both more useful features in the surgery and a lower hash rate. Second, a total of 12 surgical steps, including idle phases were identified to achieve acceptable performance in terms of recognizing surgical phases and segment boundaries. This is beneficial for establishing standard CS procedures that are highly applicable in the real world. Third, DeepSurgery not only recognized the chronological order of surgical steps and alerted surgeons to the incorrect order of steps, but also presented comparable performance to the expert panel in terms of the evaluation of surgical steps. DeepSurgery may become the paradigm for surgical management through intelligent and real-time supervision and feedback.

The finely tuned microsurgical skills of experienced cataract surgeons, achieved through their life-long development, are reflected in the steep learning curves for novice ophthalmologists [33]. CS learning curves for residents are also strongly related to feedback-based teaching guidance [34]. By providing a timely reminder for the next step and a warning for an incorrect step during CS, DeepSurgery may reduce surgical errors and guide the surgeons, especially novices, for the standardized procedures during clinical practice and CS learning. We showed that individual residents may have difficulty in self-evaluating even if explanations of the standard CS steps were provided. DeepSurgery provides an objective and standardized evaluation system for

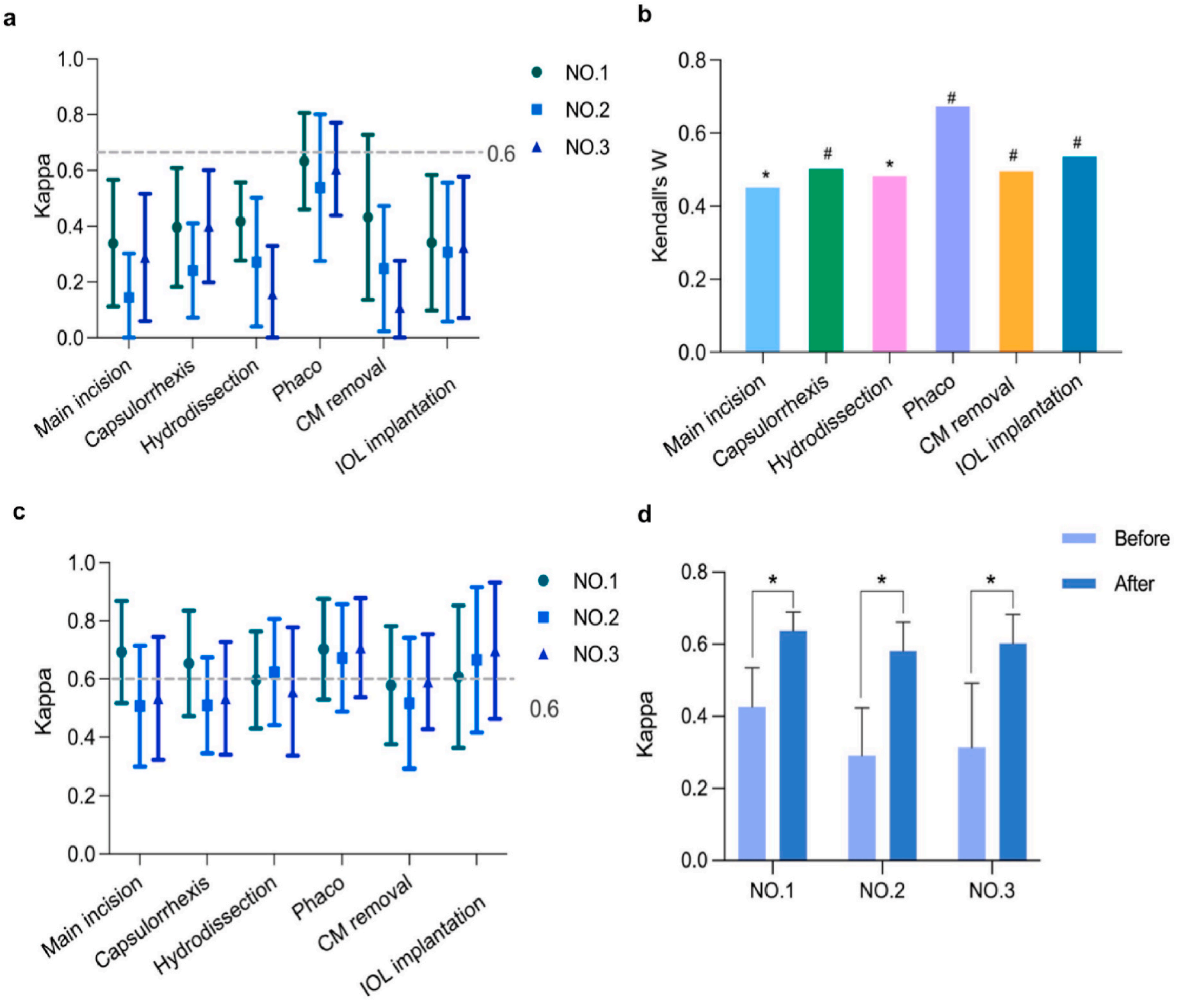

Fig. 6. The grading performance of residents in the real-time test. Six vital surgical steps were evaluated using DeepSurgery and by three residents (No. 1, No. 2, and No. 3) in real time. **a**, Except for those of the phaco step assessed by residents No. 1 and No. 3, the values for the other steps assessed by the three residents were lower than 0.6, which suggests that the residents had difficulty accurately assessing the grades of surgeons performing CS steps. **b**, The agreement among the three residents was not substantial (Kendall's W < 0.6), except for the phaco step. This finding indicates that the three residents differed in their understanding of these six CS steps. **c**, After reviewing 54 CS videos and the evaluation results given from DeepSurgery, three residents performed as well as the expert panel in phaco and IOL insertion steps. **d**, The agreement for a total of six vital steps between residents and the expert panel was significantly improved before and after reviewing the grading results of 54 CS given from DeepSurgery. The bars show 95% confidence intervals (CIs). CM: cortical material; Phaco: phacoemulsification; IOL: intraocular lens; *$p < 0.05$.

evaluating of CS steps. DeepSurgery offers a potential solution to the shortage of experienced instructors in traditional ophthalmology resident training programs. Moreover, with the development of AI and precision machinery technology, intelligent robots may become important assistants in the pursuit of improvement in surgical precision and safety. In our previous work, we proposed an autonomous robotic system for creating a self-sealing incision for use during CS [35]. DeepSurgery also has the potential to accelerate the development of such intelligent surgical robots to improve precision medicine.

Several limitations should be noted in this study. ZOC is a tertiary hospital with many complex CSs, which are usually admitted by experienced senior surgeons. Although DeepSurgery could correctly assess the nonstandard steps even when trained with a relatively small sample size, it may produce different results in different settings. In addition, our DeepSurgery method focuses on the routine steps of age-related CS, and further work is needed to evaluate different techniques used in CS. This may render DeepSurgery more applicable to ongoing training of novice surgeons as they learn new techniques and expand their repertoire. Finally, alerts are currently limited to identifying incorrect sequences of steps and grading the quality of surgery. Currently, no real-time alerts exist for maneuvers associated with increased risks of complications, such as bringing instruments too close to the posterior capsule. Further efforts are required to refine DeepSurgery in more nuanced aspects of CS to reduce the risk of complications.

5. Conclusion

DeepSurgery provides a real-time supervision and objective surgical evaluation system for routine CS, improves the quality of surgery, and reduces the demands on senior ophthalmologist trainers for surgical guidance. It may form the basis for establishing a standardized and efficient CS workflow.

Ethical approval

This study was approved by the Institutional Review Board of ZOC (No. 2021KYPJ146) and adhered to the tenets of the Declaration of Helsinki.

Sources of funding

This study was funded by Guangzhou Key Laboratory Project (No.

202002010006) and National Natural Science Foundation of China (No. 82171035).

Author statement

TW, JX and HTL contributed to the concept of the study. TW and JX designed the study. TW, XYZ and RXW did the literature search. JX, HQN and KH established the architecture of the algorithms. HTL, TW, and DYN contributed to the data collection. TW, RXW, XHW, JJC, ZZL, HC, JHW, JBL, SZL, SYY, WBC and RYL contributed to the data analysis and data interpretation. WT and JX drafted the manuscript. HTL, TW, EPL, JPOL, RYL, RXW, JX, DRL, PSY, WBC, KH, ZYX and NS critically reviewed and revised the manuscript. HTL provided research funding, coordinated the research, and oversaw the project. All the authors reviewed the manuscript for important intellectual content and approved the final manuscript.

Research registration Unique Identifying number (UIN)

Name of the registry: N/A.

Unique Identifying number or registration ID: N/A.

Hyperlink to your specific registration (must be publicly accessible and will be checked): N/A.

Guarantor

Haotian Lin.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Data sharing

Data are available on reasonable request to the corresponding author (haot.lin@hotmail.com).

Declaration of competing interest

The authors declare that there are no competing interests.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijsu.2022.106740.

References

[1] C.C. Lebares, E.V. Guvva, N.L. Ascher, P.S. O'Sullivan, H.W. Harris, E.S. Epel, Burnout and stress among US surgery residents: psychological distress and resilience, J. Am. Coll. Surg. 226 (1) (2018) 80–90. Jan.

[2] J.A. Thompson-Burdine, D.A. Telem, J.F. Waljee, E.A. Newman, D.M. Coleman, H. I. Stoll, G. Sandhu, Defining barriers and facilitators to advancement for women in academic surgery, JAMA Netw. Open 2 (8) (2019), e1910228. Aug 2.

[3] O. Ten Cate, G. Regehr, The power of subjectivity in the assessment of medical trainees, Acad. Med. 94 (3) (2019) 333–337. Mar.

[4] T. Reis, V. Lansingh, J. Ramke, J.C. Silva, S. Resnikoff, J.M. Furtado, Cataract as a cause of blindness and vision impairment in Latin America: progress made and challenges beyond 2020, Am. J. Ophthalmol. 225 (2021) 1–10. May.

[5] Trends in prevalence of blindness and distance and near vision impairment over 30 years: an analysis for the Global Burden of Disease Study, Lancet Global Health 9 (2) (2021) e130–e143. Feb.

[6] Y.C. Liu, M. Wilkins, T. Kim, B. Malyugin, J.S. Mehta, Cataracts. Lancet (London, England) 390 (10094) (2017) 600–612. Aug 5.

[7] J.T. Cox, G.B. Subburaman, B. Munoz, D.S. Friedman, R.D. Ravindran, Visual acuity outcomes after cataract surgery: high-volume versus low-volume surgeons, Ophthalmology 126 (11) (2019) 1480–1489. Nov.

[8] C.R. Garrow, K.F. Kowalewski, L. Li, M. Wagner, M.W. Schmidt, S. Engelhardt, D. A. Hashimoto, H.G. Kenngott, S. Bodenstedt, S. Speidel, B.P. Müller-Stich, F. Nickel, Machine learning for surgical phase recognition: a systematic review, Ann. Surg. 273 (4) (2021) 684–693. Apr 1.

[9] D.Z. Khan, I. Luengo, S. Barbarisi, C. Addis, L. Culshaw, N.L. Dorward, P. Haikka, A. Jain, K. Kerr, C.H. Koh, H. Layard Horsfall, W. Muirhead, P. Palmisciano, B. Vasey, D. Stoyanov, H.J. Marcus, Automated operative workflow analysis of endoscopic pituitary surgery using machine learning: development and preclinical evaluation (IDEAL stage 0), J. Neurosurg. 5 (2021) 1–8. Nov.

[10] L. Maier-Hein, S.S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G.D. Hager, P. Jannin, Surgical data science for next-generation interventions, Nat. Biomed. Eng. 1 (9) (2017) 691–696. Sep.

[11] H. Chen, C. Hu, F. Lee, C. Lin, W. Yao, L. Chen, Q. Chen, A supervised video hashing method based on a deep 3D convolutional neural network for large-scale video retrieval, Sensors 21 (9) (2021). Apr 29.

[12] N. Lu, Y. Wu, L. Feng, J. Song, Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data, IEEE J. Biomed. Health Inform. 23 (1) (2019) 314–323. Jan.

[13] L. Zhu, Y. Zhang, S. Wang, H. Yuan, S. Kwong, H.H.S. Ip, Convolutional neural network based synthesized view quality enhancement for 3D video coding, IEEE Trans. Image Process. : Public. IEEE Signal Proc. Soc. 20 (2018). Jul.

[14] F.U.M. Ullah, A. Ullah, K. Muhammad, I.U. Haq, S.W. Baik, Violence detection using spatiotemporal features with 3D convolutional neural network, Sensors 19 (11) (2019). May 30.

[15] C.Y. Zhang, Y.Y. Xiao, J.C. Lin, C.L.P. Chen, W. Liu, Y.H. Tong, 3-D deconvolutional networks for the unsupervised representation learning of human motions, IEEE Trans. Cybern. 9 (2020). Mar.

[16] M.J. Primus, D. Putzgruber-Adamitsch, M. Taschwer, B. Münzer, Y. El-Shabrawi, L. Böszörmenyi, K. Schoeffmann, Frame-Based Classification of Operation Phases in Cataract Surgery Videos, Springer International Publishing, Cham, 2018, pp. 241–253.

[17] D. Ding, W. Wang, J. Tong, X. Gao, Z. Liu, Y. Fang, Biprediction-Based Video Quality Enhancement via Learning, 17, IEEE transactions on cybernetics, 2020. Jun.

[18] K. Golnik, A. Haripriya, H. Beaver, V. Gauba, A. Lee, E. Mayorga, G. Palis, G. Saleh, Cataract surgery skill assessment, Ophthalmology 118 (10) (2011), e2092, 2094-2094.

[19] S.L. Cremers, A.N. Lora, Z.K. Ferrufino-Ponce, Global rating assessment of skills in intraocular surgery (GRASIS), Ophthalmology 112 (10) (2005) 1655–1660. Oct.

[20] S.L. Cremers, J.B. Ciolino, Z.K. Ferrufino-Ponce, B.A. Henderson, Objective assessment of skills in intraocular surgery (OASIS), Ophthalmology 112 (7) (2005) 1236–1241. Jul.

[21] V. Gulshan, L. Peng, M. Coram, M.C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P.C. Nelson, J.L. Mega, D.R. Webster, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs, JAMA 316 (22) (2016) 2402–2410. Dec 13.

[22] S. Noguchi, M. Nishio, M. Yakami, K. Nakagomi, K. Togashi, Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques, Comput. Biol. Med. 121 (2020), 103767. Jun.

[23] Z. Wang, L. Zhang, M. Zhao, Y. Wang, H. Bai, Y. Wang, C. Rui, C. Fan, J. Li, N. Li, X. Liu, Z. Wang, Y. Si, A. Feng, M. Li, Q. Zhang, Z. Yang, M. Wang, W. Wu, Y. Cao, L. Qi, X. Zeng, L. Geng, R. An, P. Li, Z. Liu, Q. Qiao, W. Zhu, W. Mo, Q. Liao, WJJocm Xu, Deep Neural Networks Offer Morphologic Classification and Diagnosis of Bacterial Vaginosis, 2020.

[24] Y. Wang, MJPmPaijdttaoptm Wang, Physics bojotIAoB, Selecting Proper Combination of mpMRI Sequences for Prostate Cancer Classification Using Multi-Input Convolutional Neuronal Network, 80, 2020, pp. 92–100.

[25] D.L. Kennedy, H.I. Kemp, D. Ridout, D. Yarnitsky, A.S.C. Rice, Reliability of conditioned pain modulation: a systematic review, Pain 157 (11) (2016) 2410–2419. Nov.

[26] M. Winters, E.W.P. Bakker, M.H. Moen, C.C. Barten, R. Teeuwen, A. Weir, Medial tibial stress syndrome can be diagnosed reliably using history and physical examination, Br. J. Sports Med. 52 (19) (2018) 1267–1272. Oct.

[27] E. Duregon, A. Cassenti, A. Pittaro, L. Ventura, R. Senetta, R. Rudà, PJN-o Cassoni, Better See to Better Agree: Phosphohistone H3 Increases Interobserver Agreement in Mitotic Count for Meningioma Grading and Imposes New Specific Thresholds, 17, 2015, pp. 663–669, 5.

[28] O. Zisimopoulos, E. Flouty, M. Stacey, S. Muscroft, P. Giataganas, J. Nehme, A. Chow, D. Stoyanov, Can surgical simulation be used to train detection and classification of neural networks? Healthcare Technol. Lett. 4 (5) (2017) 216–222. Oct.

[29] H. Al Hajj, M. Lamard, P.H. Conze, S. Roychowdhury, X. Hu, G. Maršalkaitė, O. Zisimopoulos, M.A. Dedmari, F. Zhao, J. Prellberg, M. Sahu, A. Galdran, T. Araújo, D.M. Vo, C. Panda, N. Dahiya, S. Kondo, Z. Bian, A. Vahdat, J. Bialopetravičius, E. Flouty, C. Qiu, S. Dill, A. Mukhopadhyay, P. Costa, G. Aresta, S. Ramamurthy, S.W. Lee, A. Campilho, S. Zachow, S. Xia, S. Conjeti, D. Stoyanov, J. Armaitis, P.A. Heng, W.G. Macready, B. Cochener, G. Quellec, CATARACTS: challenge on automatic tool annotation for cataRACT surgery, Med. Image Anal. 52 (2019) 24–41. Feb.

[30] S. Morita, H. Tabuchi, H. Masumoto, H. Tanabe, N. Kamiura, Real-time surgical problem detection and instrument tracking in cataract surgery, J. Clin. Med. 9 (12) (2020). Nov 30.

[31] F. Yu, G. Silva Croso, T.S. Kim, Z. Song, F. Parker, G.D. Hager, A. Reiter, S. S. Vedula, H. Ali, S. Sikder, Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques, JAMA Netw. Open 2 (4) (2019), e191860. Apr 5.

[32] G. Quellec, M. Lamard, B. Cochener, G. Cazuguel, Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials, IEEE Trans. Med. Imag. 34 (4) (2015) 877–887. Apr.

[33] S. Gupta, A. Haripriya, S.A. Vardhan, T. Ravilla, R.D. Ravindran, Residents' learning curve for manual small-incision cataract surgery at aravind eye hospital, India, Ophthalmology 125 (11) (2018) 1692–1699. Nov.

[34] K. Kaplowitz, M. Yazdanie, A. Abazari, A review of teaching methods and outcomes of resident phacoemulsification, Surv. Ophthalmol. 63 (2) (2018) 257–267. Mar-Apr.

[35] J. Xia, S.J. Bergunder, D. Lin, Y. Yan, S. Lin, M. Ali Nasseri, M. Zhou, H. Lin, K. Huang, Microscope-guided autonomous clear corneal incision, in: 2020 IEEE International Conference on Robotics and Automation, ICRA, Institute of Electrical and Electronics Engineers Inc., Paris, France, 2020, pp. 3867–3873. May 31, 2020 - August 31, 2020: 2020.