



Efficient Spatiotemporal Learning of Microscopic Video for Augmented Reality-Guided Phacoemulsification Cataract Surgery

Puxun Tu¹ , Hongfei Ye², Jeff Young³, Meng Xie², Ce Zheng² ,
and Xiaojun Chen^{1,4}

¹ Institute of Biomedical Manufacturing and Life Quality Engineering, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China
xiaojunchen@sjtu.edu.cn

² Department of Ophthalmology, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China
zhengce@xinhuamed.com.cn

³ Department of Bioengineering, University of Texas at Dallas, Richardson, USA

⁴ Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China

Abstract. Phacoemulsification cataract surgery (PCS) is typically performed under a surgical microscope and adhering to standard procedures. The success of this surgery depends heavily on the seniority and experience of the ophthalmologist performing it. In this study, we developed an augmented reality (AR) guidance system to enhance the intraoperative skills of ophthalmologists by proposing a two-stage spatiotemporal learning network for surgical microscope video recognition. In the first stage, we designed a multi-task network that recognizes surgical phases and segments the limbus region to extract limbus-focused spatial features. In the second stage, we developed a temporal pyramid-based spatiotemporal feature aggregation (TP-SFA) module that uses causal and dilated temporal convolution for smooth and online surgical phase recognition. To provide phase-specific AR guidance, we designed several intraoperative visual cues based on the parameters of the fitted limbus ellipse and the recognized surgical phase. The comparison experiments results indicate that our method outperforms several strong baselines in surgical phase recognition. Furthermore, ablation experiments show the positive effects of the multi-task feature extractor and TP-SFA module. Our developed system has the potential for clinical application in PCS to provide real-time intraoperative AR guidance.

Keywords: Cataract surgery · Augmented reality · Surgical phase recognition · Spatiotemporal learning

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43990-2_64.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14226, pp. 682–692, 2023.
https://doi.org/10.1007/978-3-031-43990-2_64

1 Introduction

Cataract is the leading cause of blindness worldwide, and cataract surgery is one of the most common operations in health care. Among different cataract surgery techniques, phacoemulsification cataract surgery (PCS) is the standard of care [5, 23]. PCS consists of the manual opening of the crystalline lens anterior capsule with forceps, removal of the opacified lens, and implantation of an intraocular lens (IOL) in the remaining capsular envelope to restore visual function.

Phacoemulsification needs microsurgical skills, which depend on numerous variables, including the amount of practice, inherent manual dexterity, and previous experience. Statistical analyses have demonstrated significant differences in completion and complication rates among ophthalmologists, with variations observed based on factors such as seniority and experience [14]. During phacoemulsification, a surgical microscope with an integrated video camera is routinely used, providing rich spatiotemporal information [15]. This information presents an excellent opportunity to develop surgical video recognition methods to extract valuable intraoperative information and overlay it on a 2D/3D screen or the microscopic eyepiece, thereby creating an augmented reality (AR) scene.

To bridge the experience gap among ophthalmologists, several intraoperative AR-guided systems have been developed. Zhai et al. [25] used two convolutional neural networks (CNNs) to segment the limbus and track the eye’s rotation, and subsequently developed an intraoperative guide system for positioning and aligning the IOL. In [16], a multi-task CNN was designed to locate the pupil and recognize the surgical phase in each frame of the surgical microscope video. Nespole et al. [17] utilized a deep CNN-based method for processing surgical videos, allowing for detecting surgical instruments and tissue boundaries to guide the ophthalmic surgery. Despite the potential of AR-guided phacoemulsification systems, limitations still hinder their clinical implementation. Firstly, the current systems process surgical videos in a frame-wise manner, enabling real-time processing but leading to lost temporal information and incoherent surgical scene recognition. Secondly, the overlaid information during surgery is not categorized by surgical phase, causing visual redundancy for ophthalmologists.

Advancements in video spatiotemporal learning, particularly in surgical phase recognition, present a promising opportunity to switch AR scenes to the current surgical phase automatically. Early attempts used a 2D CNN to extract spatial features to predict each video frame’s surgical phase [18, 22]. However, the lack of temporal information leads to unsatisfactory recognition accuracy. Other studies [11, 24] use spatial feature maps of neighboring frames, extracted from CNNs, as an input to Gated Recurrent Unit (GRU) [2] or Long Short-Term Memory (LSTM) [10] to model the temporal dependencies and predict the surgical phase. However, these methods suffer from limited temporal reception field and non-parallel, slow inference. Recent studies focus on modeling long-range temporal relationships. Czempiel et al. [3] introduced a multi-stage temporal convolutional network (TCN) for surgical phase recognition, leveraging causal and dilated convolutions to enable global reception field and online deployment. The current state-of-the-art methods utilize transformer-based models for aggre-

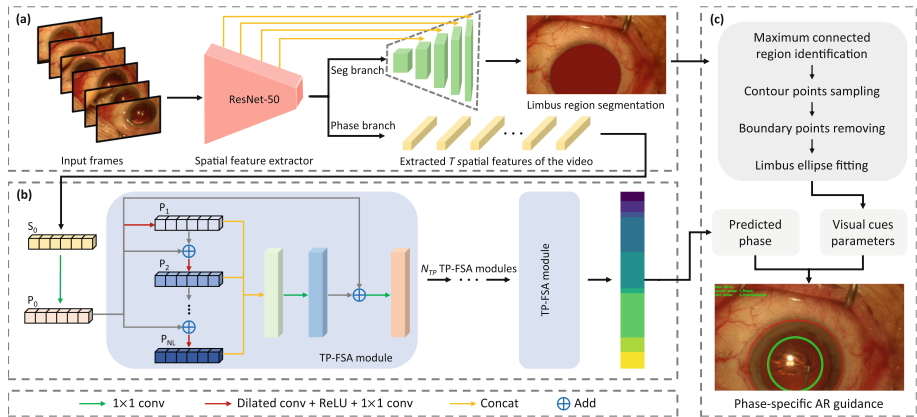


Fig. 1. Overview of our AR-guided system for PCS. (a) Limbus region-focused spatial feature extraction. (b) TP-SFA module-based spatiotemporal aggregation. (c) Phase-specific intraoperative AR guidance.

gating spatial [4, 27] and spatiotemporal features [8, 12]. However, focusing solely on global features may result in losing important local temporal dependencies and lead to inaccurate recognition of challenging frames.

In this study, we developed a novel intraoperative AR-guide system for PCS. Our contributions are two-fold: (1) We propose an efficient spatiotemporal network for surgical microscope video recognition, consisting of two stages: a multi-task learning stage for limbus segmentation and spatial feature extraction, and a temporal pyramid-based spatiotemporal feature aggregation (TP-SFA) module for online surgical phase recognition. (2) We use the limbus and surgical phase information to design phase-specific visual cues that offer real-time intraoperative AR guidance while avoiding distracting the ophthalmologist’s attention.

2 Methods

Figure 1 presents an overview of our developed AR-guided system for PCS, which acquires intraoperative video streams from microscope and processes them using the proposed two-stage spatiotemporal network to obtain limbus and surgical phase information. Parameters of intraoperative visual cues are computed using the fitted limbus elliptic parameters and updated according to the recognized surgical phase, providing automatic AR scene switching for ophthalmologists.

2.1 Spatiotemporal Network for Microscopic Video Recognition

Limbus Segmentation and Spatial Feature Extraction. We observe that the region within the limbus displays distinguishable appearances at different phases in surgical microscope videos, whereas other regions like the sclera exhibit

similar appearances. We argue that using a limbus region-focused spatial feature extraction network can improve spatiotemporal aggregation. This led us to develop a multi-task network for limbus segmentation and phase recognition in the first stage. As shown in Fig. 1(a), we employ ResNet-50 [9] as the shared backbone to extract spatial features, which are then fed into both the surgical phase recognition and limbus segmentation branches.

The surgical phase recognition branch involves a fully connected layer that is directly connected to the global average pooling layer, followed by a softmax layer. To train this branch, we use cross-entropy loss, which is defined as

$$L_{phase} = -(\sum_{s=1}^{N_s} g_s \log p_s) / N_s, \quad (1)$$

where g_s is the ground truth binary indicator of phase s , p_s is the probability of the input frame belonging to phase s .

The limbus segmentation branch incorporates a decoder with upsampling and concatenation, resembling the U-net [20] architecture. To train this branch, we employ a hybrid loss of cross-entropy and dice, which is defined as

$$L_{seg} = -\frac{1}{N \times C} \sum_{c=1}^C \sum_{i=1}^N y_i^c \log p_i^c + \alpha \left(1 - \frac{2 \sum_{c=1}^C \sum_{i=1}^N y_i^c p_i^c}{\sum_{c=1}^C \sum_{i=1}^N y_i^c + \sum_{c=1}^C \sum_{i=1}^N p_i^c} \right), \quad (2)$$

where y_i^c and p_i^c are the pixel-level ground truth and prediction result respectively, α is a hyper-parameter to balance the loss. The final loss function for training the first stage is defined as

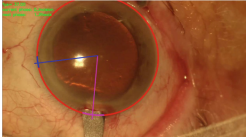
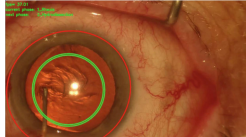
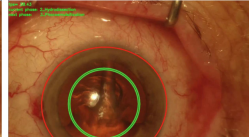
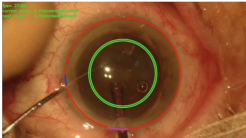
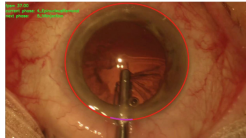
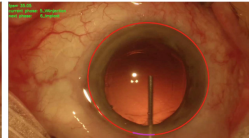
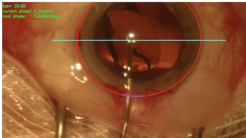
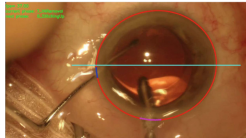
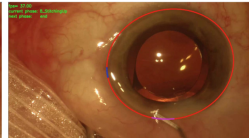
$$L_{spatial} = L_{phase} + \beta L_{seg}, \quad (3)$$

where β is a hyper-parameter to balance the loss. After training the first stage, we obtain the spatial feature $s_t \in \mathbb{R}^{2048}$ for frame t by outputting the average pooling layer of the spatial feature extractor.

Spatiotemporal Features Aggregation. We argue that the surgical phase recognition method used for intraoperative AR should fulfill the following requirements: 1) online recognition for real-time intraoperative guidance, and 2) sufficient stability to avoid distracting ophthalmologists with incorrect phase recognition. As shown in Fig. 1(b), we employ our proposed TP-SFA module in the second stage, which uses multi-scale, causal, and dilated temporal convolutions to model the temporal relationships. We denote the input spatial feature as $S_0 \in \mathbb{R}^{2048 \times T}$, where T is the sequence length of the video. The first layer of the TP-SFA module is a 1×1 convolutional layer that reduces the dimension of S_0 and outputs $P_0 \in \mathbb{R}^{32 \times T}$. To obtain temporal features with different reception fields, we apply N_L dilated layers with different dilation factors on P_0 . Layer k consists of a dilated convolution with a dilation factor of 2^k , followed by a ReLU activation and a 1×1 convolution. This can be described as

$$P_k = W_2 * \text{ReLU}(W_1 * (P_{k-1} + P_0) + b_1) + b_2, \quad (4)$$

Table 1. AR scene at different phases with different combinations of visual cues for PCS. The color of visual cues in both the text and AR scene figure is consistent.

AR Scene			
Phase	Incision	Rhexis	Hydrodissection
Visual cues	FLE+PIC+SIC+IGL	FLE+PIC+RR	FLE+PIC+RR
AR Scene			
Phase	Phacoemulsification	Epinucleus removal	VA injection
Visual cues	FLE+PIC+SIC+RR	FLE+PIC	FLE+PIC
AR Scene			
Phase	Implant setting-up	VA removal	Stitching up
Visual cues	FLE+PIC+IRL	FLE+PIC+SIC+IRL	FLE+PIC+SIC

where $*$ denotes the convolutional operator, W_1 is the dilated convolution weights, W_2 is the weights of the 1×1 convolution, and b_1 and b_2 are bias vectors. Finally, we concatenate all $P_k (k = 1, \dots, N_L)$ together over the temporal dimension, followed by a 1×1 convolution, a residual connection with P_0 and another 1×1 convolution to adjust the output dimension. This can be described as

$$S_1 = W_4 * (W_3 * \text{concat}(P_1, \dots, P_{N_L}) + b_3 + P_0) + b_4, \quad (5)$$

where W_3 and W_4 are the weights of the 1×1 convolution, and b_3 and b_4 are bias vectors. Inspired by [3, 7], we connect N_{tp} TP-SFA modules together and use a weighted cross-entropy loss to train the second stage. This can be described as

$$L_{temp} = -\frac{1}{N_{tp} \times N_s} \sum_{n=1}^{N_{tp}} \sum_{s=1}^{N_s} w_s g_s^n \log p_s^n, \quad (6)$$

where w_s is the weight and is inversely proportional to the surgical phase frequencies [6].

2.2 Phase-Specific AR Guidance in PCS

The proposed spatiotemporal learning network enables real-time limbus segmentation and surgical phase, facilitating the development of our intraoperative AR

guidance system for PCS. The limbus contour can be fitted as an ellipse [25, 26]. To accomplish this, we follow the steps shown in Fig. 1(c), including: 1) identifying the maximum connected region to remove possible mis-segmented regions; 2) extracting the contour of the maximum connected region and sampling the contour points; 3) removing contour points near the video boundaries; 4) fitting the remaining contour points to an ellipse and outputting the length and rotation of the long and short axes of the ellipse.

We segment PCS into nine phases [19]: incision, rhexis, hydrodissection, phacoemulsification, epinucleus removal, viscous agent (VA) injection, implant setting-up, VA removal, and stitching up. For intraoperative AR guidance, we designed six visual cues, including: 1) fitted limbus ellipse (FLE), extracted from the segmentation results; 2) primary incision curve (PIC), defined as an arc with a length equal to the maximum diameter of the primary incision knife; 3) secondary incision curve (SIC), defined as an arc with a length equal to the maximum diameter of the secondary incision knife; 4) incision guide lines (IGL), with included angles of 95^{circ} for BIC and 173^{circ} for SIC, respectively, relative to the reference line; 5) rhexis region (RR), with a diameter equal to half the length of the long axis of the fitted ellipse; and 6) implant reference line (IRL), defined by a horizontal line. Table 1 lists different combinations of intraoperative visual cues that are automatically updated according to the recognized surgical phase.

3 Experiments and Results

3.1 Dataset and Implementation Details

Dataset. We evaluate our methods on CATARACTS [1], a publicly available dataset for cataract surgery. It contains 50 videos with a frame rate of 30 frames-per-second (fps) and a total duration of over nine hours. All videos have been subsampled to 1 fps. Each frame has a resolution of 1920×1080 pixels and has been annotated in nineteen surgical steps. For the sake of intraoperative guidance convenience, we have reorganized the nineteen fine-grained surgical steps into nine standard phases [19]. Additionally, the limbus region of each frame has been manually delineated by two non-M.D. experts. The dataset is split into 25 cases for training, 5 cases for validation, and 20 cases for test, following [1, 12].

Implementation Details. Our network was implemented in PyTorch using two NVIDIA GeForce GTX 3090 GPUs. We initialized the ResNet-50 backbone with weights trained on the ImageNet [21] and implemented random horizontal flip, random crop, random rotation ($\pm 20^{circ}$), and color jitter for data augmentation. The first stage was trained for 100 epochs using Adam optimizer with a learning rate of $5e-5$ for the backbone and $5e-4$ for the fully connected layer and decoder. The second stage was trained for 50 epochs using Adam optimizer with a learning rate of $1e-4$. For hyper-parameters, we set $\alpha = 0.6$, $\beta = 0.5$ and $N_L=8$.

Table 2. Quantitative comparison results with strong baselines in online surgical phase recognition. Each metric is reported as mean (%) and standard deviation (\pm).

Methods	Accuracy	Precision	Recall	Jaccard
ResNet-50 [9]	81.1 \pm 6.4	78.4 \pm 8.0	77.9 \pm 7.1	63.7 \pm 9.4
SV-RCNet [11]	84.8 \pm 6.2	81.1 \pm 7.3	81.8 \pm 6.8	69.1 \pm 8.5
TeCNO [3]	86.1 \pm 5.5	81.6 \pm 6.7	83.5 \pm 6.3	70.8 \pm 7.1
Trans-SVNet [8]	86.8 \pm 5.9	82.7 \pm 7.1	83.6 \pm 6.5	71.6 \pm 7.0
Ours	87.9 \pm 5.4	83.7 \pm 6.6	84.5 \pm 5.8	73.3 \pm 6.3

Table 3. Phase recognition and segmentation results of multi-task and single task.

Method	Accuracy		Dice
	First stage	Second stage	
Single phase task	81.1 \pm 6.4	85.6 \pm 5.9	—
Single segmentation task	—	—	94.0 \pm 2.3
Phase+Segmentation	82.6 \pm 5.8	87.9 \pm 5.4	94.6 \pm 2.7

3.2 Comparisons with Strong Baselines

We compare our method with several strong baselines in surgical phase recognition: 1) ResNet-50 [9], a deep residual network that predicts surgical phase in a frame-wise manner; 2) SV-RCNet [11], an end-to-end architecture that utilizes LSTM to learn temporal features; 3) TeCNO [3], a multi-stage temporal convolutional network that models global temporal features; and 4) Trans-SVNet [8], a transformer-based spatiotemporal features aggregation network. Note that we only included comparison methods that support online surgical phase recognition and excluded those that do not. For fair comparison, we use the proposed multi-task learning network in the first stage as the spatial feature extraction backbone for all methods except ResNet-50 [9]. The quantitative comparison results are listed in Table 2, which indicates that our method achieved the best performance among all compared methods. We show the quantitative results with color-coded ribbons in Fig. 2(a). The results indicate that our method can produce a smoother phase prediction compared to ResNet-50 [9] and SV-RCNet [11]. Additionally, our approach surpasses the performance of global feature aggregation-based methods [3, 8] in challenging local frames.

3.3 Ablation Study

Effect of Multi-task Feature Extractor We performed experiments to evaluate the effect of the multi-task feature extractor. ResNet-50 [9] served as the backbone for all methods. We compared the accuracy of the first stage and the second stage for the single phase recognition task with that of the multi-task

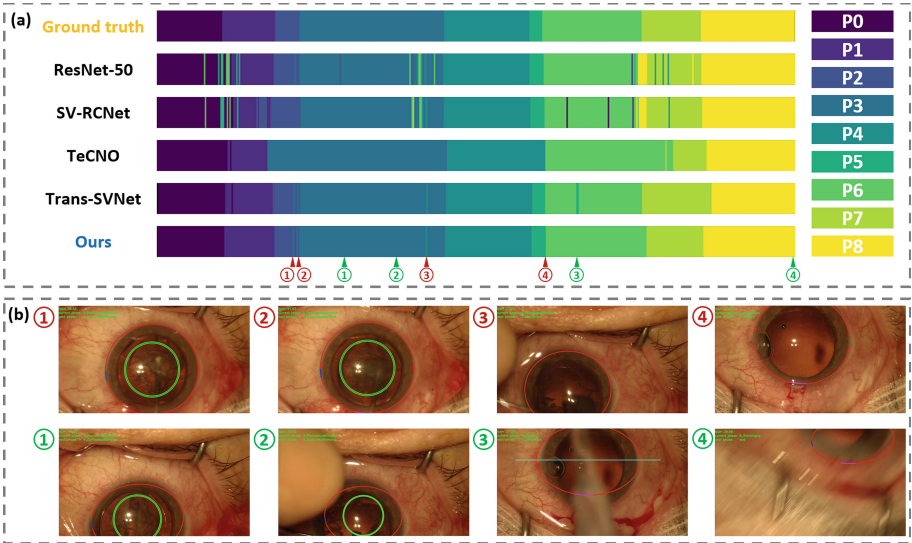


Fig. 2. (a) The comparison results with color-coded ribbons. (b) Some typical failed AR scenes. Markers in red and green represent failed scenes caused by mis-recognition of the surgical phase and mis-segmentation of the limbus, respectively. (Color figure online)

Table 4. The effect of different number of TP-SFA modules.

Table 5. The effect of different combinations of TP-SFA and causal TCN.

N_{TP}	Accuracy	First module		Second module		Accuracy
		TP-SFA	TCN	TP-SFA	TCN	
1	85.7 ± 6.9	✓		✓		87.9 ± 5.4
2	87.9 ± 5.4		✓		✓	86.0 ± 5.7
3	87.6 ± 5.8	✓			✓	87.2 ± 5.2
4	87.3 ± 5.2		✓	✓		86.4 ± 6.3

approach. Moreover, we compared the segmentation Dice score of the single segmentation task with that of the multi-task approach. Table 3 shows the results, indicating that the multi-task feature extractor enhances both segmentation and phase recognition performance compared to the single-task approach.

Effect of the TP-SFA Module We evaluate the number of the connected TP-SFA modules. The quantitative results are listed in Table 4, which indicates that the second stage achieves the best accuracy when two TP-SFA modules are used. Furthermore, we explore different combinations of the TP-SFA module and a typical causal TCN module [13] and present the results in Table 5. Results show that two connected TP-SFA modules achieve the best performance.

3.4 AR Guidance Evaluation

Our method achieves real-time intraoperative processing at a speed of 36 fps. This makes it suitable for meeting the demands of online intraoperative AR guidance, as the acquired microscope video stream has a speed of 30 fps. We show some typical failed scenes in Fig. 2(b). Failed intraoperative AR guidance can result from both mis-recognition of surgical phases and mis-segmentation of the limbus. Mis-recognition of surgical phase may introduce continuous AR scene switching problem and distract the ophthalmologist's attention.

4 Conclusion

We proposed a two-stage spatiotemporal network for online microscope video recognition. Furthermore, we developed a phase-specific intraoperative AR guidance system for PCS. Our developed system has the potential for clinical applications to enhance ophthalmologists' intraoperative skills.

Acknowledgements. This work was supported by grants from the National Natural Science Foundation of China (81971709; M-0019; 82011530141), the Foundation of Science and Technology Commission of Shanghai Municipality (20490740700; 22Y11911700), Shanghai Jiao Tong University Foundation on Medical and Technological Joint Science Research (YG2021ZD21; YG2021QN72; YG2022QN056; YG2023ZD19; YG2023ZD15), Hospital Funded Clinical Research, Xinhua Hospital Affiliated to Shanghai Jiao Tong University School of Medicine (21XJMR02), and the Funding of Xiamen Science and Technology Bureau (No. 3502Z20221012).

References

1. Al Hajj, H., et al.: CATARACTS: challenge on automatic tool annotation for cataract surgery. *Med. Image Anal.* **52**, 24–41 (2019)
2. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint [arXiv:1412.3555](https://arxiv.org/abs/1412.3555) (2014)
3. Czempiel, T., et al.: TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. In: Martel, A.L., et al. (eds.) *MICCAI 2020. LNCS*, vol. 12263, pp. 343–352. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_33
4. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: OperA: attention-regularized transformers for surgical phase recognition. In: de Bruijne, M., et al. (eds.) *MICCAI 2021. LNCS*, vol. 12904, pp. 604–614. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_58
5. Day, A.C., Gore, D.M., Bunce, C., Evans, J.R.: Laser-assisted cataract surgery versus standard ultrasound phacoemulsification cataract surgery. *Cochrane Database of Systematic Reviews* (7) (2016)
6. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2650–2658 (2015)

7. Farha, Y.A., Gall, J.: MS-TCN: multi-stage temporal convolutional network for action segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584 (2019)
8. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.-A.: Trans-SVNet: accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: de Bruijne, M., et al. (eds.) *MICCAI 2021*. LNCS, vol. 12904, pp. 593–603. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_57
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**(8), 1735–1780 (1997)
11. Jin, Y., et al.: SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE Trans. Med. Imaging* **37**(5), 1114–1126 (2017)
12. Jin, Y., Long, Y., Gao, X., Stoyanov, D., Dou, Q., Heng, P.A.: Trans-SVNet: hybrid embedding aggregation transformer for surgical workflow analysis. *Int. J. Comput. Assist. Radiol. Surg.* **17**(12), 2193–2202 (2022)
13. Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: a unified approach to action segmentation. In: Hua, G., Jégou, H. (eds.) *ECCV 2016*. LNCS, vol. 9915, pp. 47–54. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_7
14. Lee, J.S., Hou, C.H., Lin, K.K.: Surgical results of phacoemulsification performed by residents: a time-trend analysis in a teaching hospital from 2005 to 2021. *J. Ophthalmol.* **2022** (2022)
15. Ma, L., Fei, B.: Comprehensive review of surgical microscopes: technology development and medical applications. *J. Biomed. Opt.* **26**(1), 010901–010901 (2021)
16. Nespolo, R.G., Yi, D., Cole, E., Valikodath, N., Luciano, C., Leiderman, Y.I.: Evaluation of artificial intelligence-based intraoperative guidance tools for phacoemulsification cataract surgery. *JAMA Ophthalmol.* **140**(2), 170–177 (2022)
17. Nespolo, R.G., Yi, D., Cole, E., Wang, D., Warren, A., Leiderman, Y.I.: Feature tracking and segmentation in real time via deep learning in vitreoretinal surgery—a platform for artificial intelligence-mediated surgical guidance. *Ophthalmol. Retina* **7**(3), 236–242 (2022)
18. Primus, M.J.: Frame-based classification of operation phases in cataract surgery videos. In: Schoeffmann, K., et al. (eds.) *MMM 2018*. LNCS, vol. 10704, pp. 241–253. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73603-7_20
19. Quellec, G., Lamard, M., Cochener, B., Cazuguel, G.: Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials. *IEEE Trans. Med. Imaging* **34**(4), 877–887 (2014)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Russakovsky, O.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015)
22. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **36**(1), 86–97 (2016)
23. Wang, W., et al.: Cataract surgical rate and socioeconomics: a global study. *Invest. Ophthalmol. Vis. Sci.* **57**(14), 5872–5881 (2016)

24. Yi, F., Yang, Y., Jiang, T.: Not end-to-end: explore multi-stage architecture for online surgical phase recognition. In: Proceedings of the Asian Conference on Computer Vision, pp. 2613–2628 (2022)
25. Zhai, Y., et al.: Computer-aided intraoperative toric intraocular lens positioning and alignment during cataract surgery. *IEEE J. Biomed. Health Inform.* **25**(10), 3921–3932 (2021)
26. Zhao, W., Zhang, Z., Wang, Z., Guo, Y., Xie, J., Xu, X.: ECLNet: center localization of eye structures based on adaptive gaussian ellipse heatmap. *Comput. Biol. Med.* **153**, 106485 (2023)
27. Zou, X., Liu, W., Wang, J., Tao, R., Zheng, G.: ARST: auto-regressive surgical transformer for phase recognition from laparoscopic videos. *Comput. Meth. Biomech. Biomed. Eng. Imaging Visual.* **11**, 1012–1018 (2022)