

Cataract-1K: Cataract Surgery Dataset for Scene Segmentation, Phase Recognition, and Irregularity Detection

Negin Ghamsarian¹, Yosuf El-Shabrawi³, Sahar Nasirihaghghi², Doris Putzgruber-Adamitsch³, Martin Zinkernagel⁴, Sebastian Wolf⁴, Klaus Schoeffmann^{2*}, and Raphael Sznitman¹

¹Center for Artificial Intelligence in Medicine (CAIM), Department of Medicine, University of Bern, Switzerland

²Department of Information Technology, University of Klagenfurt, Austria

³Department of Ophthalmology, Klinikum Klagenfurt, Austria

⁴Department of Ophthalmology, Inselspital, Bern, Switzerland

*corresponding author: (ks@itec.aau.at)

ABSTRACT

In recent years, the landscape of computer-assisted interventions and post-operative surgical video analysis has been dramatically reshaped by deep-learning techniques, resulting in significant advancements in surgeons' skills, operation room management, and overall surgical outcomes. However, the progression of deep-learning-powered surgical technologies is profoundly reliant on large-scale datasets and annotations. Particularly, surgical scene understanding and phase recognition stand as pivotal pillars within the realm of computer-assisted surgery and post-operative assessment of cataract surgery videos. In this context, we present the largest cataract surgery video dataset that addresses diverse requisites for constructing computerized surgical workflow analysis and detecting post-operative irregularities in cataract surgery. We validate the quality of annotations by benchmarking the performance of several state-of-the-art neural network architectures for phase recognition and surgical scene segmentation. Besides, we initiate the research on domain adaptation for instrument segmentation in cataract surgery by evaluating cross-domain instrument segmentation performance in cataract surgery videos. The dataset and annotations will be publicly available upon acceptance of the paper.

Background & Summary

Following the technological advancements in surgery, operation rooms are evolving into intelligent environments. Context-aware systems (CAS) are emerging as pivotal components of this evolution, empowered to advance pre-operative surgical planning^{1–3}, automate skill assessment^{4–8}, support operation room planning^{9–11}, and interpret the surgical context comprehensively. By enabling real-time alerts and offering decision-making support, these systems prove invaluable, especially but not only for less-experienced surgeons. Their capabilities extend to the automatic analysis of surgical videos, encompassing functions like indexing, documentation, and generating post-operative reports¹². The ever-increasing demand for such automatic systems has sparked machine-learning-based approaches to surgical video analysis.

Cataract Surgery, renowned as the most commonly conducted ophthalmic surgical procedure and one of the most demanding surgeries worldwide, is a major operation where deep learning can be of great benefit. Cataract, characterized by the opacification of the eye's natural lens, is often attributed to aging and leads to impaired visual acuity, reduced brightness, visual distortion, double vision, and color perception degradation. Globally, cataracts stand as the primary cause of blindness¹³. Owing to the aging demographic and increased lifespans, the World Health Organization forecasts a surge in cataract-related blindness cases, estimating the number to reach 40 million by the year 2025¹³. This prevalent disease can be remedied through cataract surgery involving the substitution of the eye's natural lens with a synthetic counterpart known as an intraocular lens (IOL). Advancements in technology have driven the evolution of cataract surgery techniques. This evolution spans from intracapsular cataract extraction (ICCE) in the 1960s and 1970s to extracapsular cataract extraction (ECCE) in the 1980s and 1990s. Today, the primary method involves sutureless small-incision phacoemulsification surgery with an injectable intraocular lens (IOL) implantation¹. Due to the widespread occurrence of cataract surgery and its substantial influence on patients' quality of life, a significant focus has been directed towards the analysis of cataract surgery content using deep learning methodologies.

¹Throughout this paper, the term "Cataract Surgery" is synonymous with "Phacoemulsification Cataract Surgery."

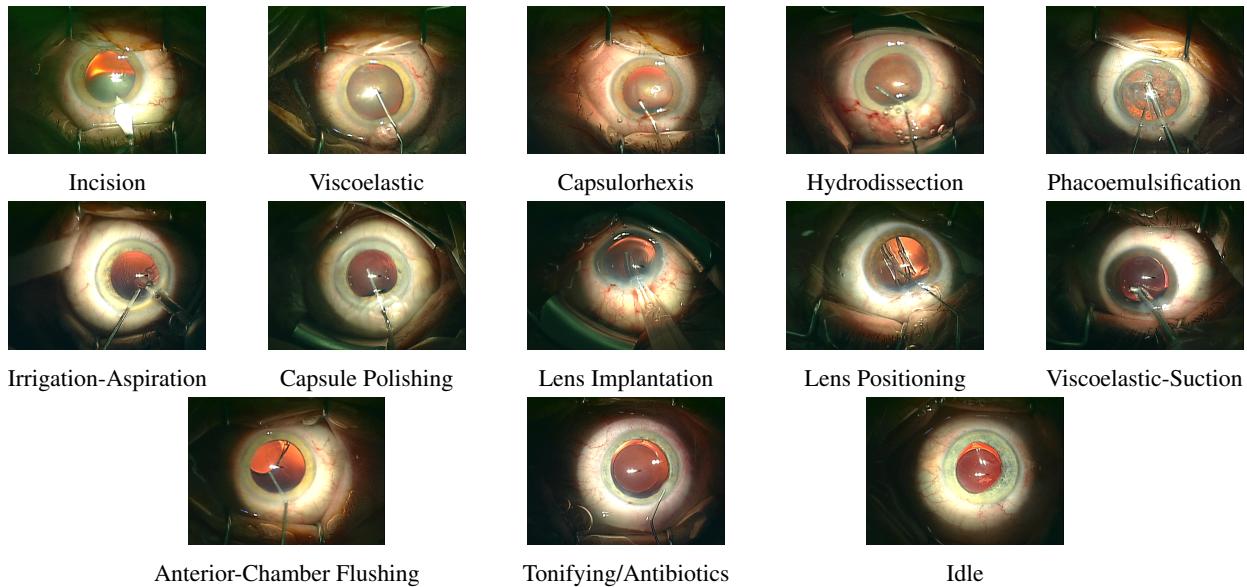


Figure 1. Sample frames from different phases in a regular cataract surgery.

over the past decade. In particular, Surgical phase recognition and scene segmentation are joint building blocks in various applications related to cataract surgery video analysis¹². These applications include but are not limited to relevance detection¹⁴, relevance-based compression¹⁵, irregularity detection^{16,17}, and surgical outcome prediction¹⁸. The current public datasets for cataract surgery either provide annotations for a particular sub-task such as instrument recognition¹⁹, scene and relevant anatomical structure segmentation^{14,20-22}, or offer small multi-task datasets targeting specific problems such as intraocular lens (IOL) irregularity detection¹⁶. As a result of the lack of a comprehensive dataset, there exists a considerable gap in exploring deep-learning-based approaches and frameworks to enhance cataract surgery outcomes. To facilitate the development of such systems and models, there is a compelling need for large-scale datasets that encompass multi-task annotations.

This paper introduces the largest cataract surgery video dataset, including 1000 videos of cataract surgery recorded in Klinikum Klagenfurt, Austria, between 2021 and 2023. We provide large-scale ground-truth annotations for the semantic segmentation of different instruments and relevant anatomical structures, as well as surgical phases. Besides, the dataset features two subsets for major irregularities in cataract surgery, which affect surgical workflow, including intraocular lens (IOL) rotation, and pupil contraction in cataract surgery. Together, these 1000 videos, annotated datasets, and irregularity subsets form a complete dataset to empower computer-assisted interventions (CAI) in cataract surgery.

Methods

Cataract-1K Dataset Description

The Cataract-1K dataset consists of 1000 videos of cataract surgeries performed in the eye clinic of Klinikum Klagenfurt from 2021 to 2023². From these videos, we provide surgical phase annotations for 56 regular videos and relevant anatomical plus instrument pixel-level annotations for 2256 frames out of 30 cataract surgery videos. Furthermore, we provide a small subset of surgeries with two major irregularities, including "pupil reaction" and "IOL rotation," to support further research on irregularity detection in cataract surgery. Except for the annotated videos and images, the remaining videos in the Cataract-1K dataset are encoded with a temporal resolution of 25 fps and a spatial resolution of 512 × 324. Besides, we assess the surgeons' skills by considering the cumulative count of their completed surgeries, which spans from 1,000 to over 40000 procedures in the Cataract-1K dataset. We delineate the challenges and annotation procedures for each subset in the following paragraphs.

Phase recognition dataset. Crafting an approach to detect and classify significant phases within these videos, considering frame-by-frame temporal details, presents considerable challenges due to several factors:

- As illustrated in Figure 1, instruments, which play a fundamental role in distinguishing between relevant phases, share a substantial resemblance in certain phases, leading to a narrow variation between different classes in a trained classification model.

²The dataset will be publicly released in Synapse upon paper acceptance. For anonymizing purposes, however, our dataset is temporarily accessible via Figshare.

Table 1. Visualizations of phase annotations for 56 normal cataract surgeries. The durations of the videos are different and normalized for better visualization.

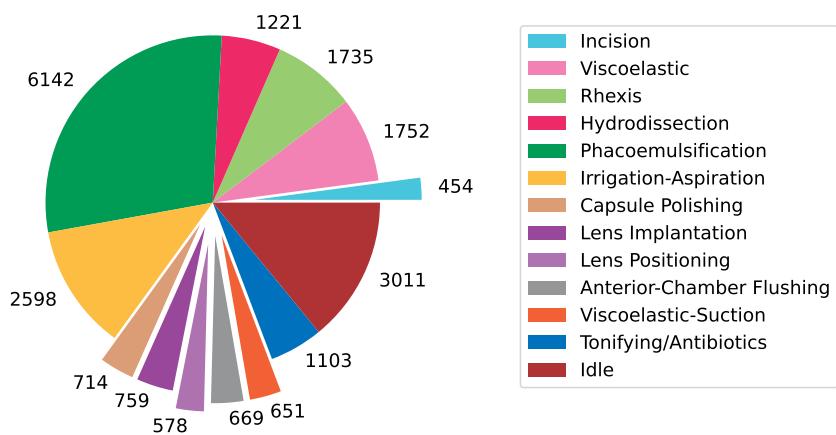
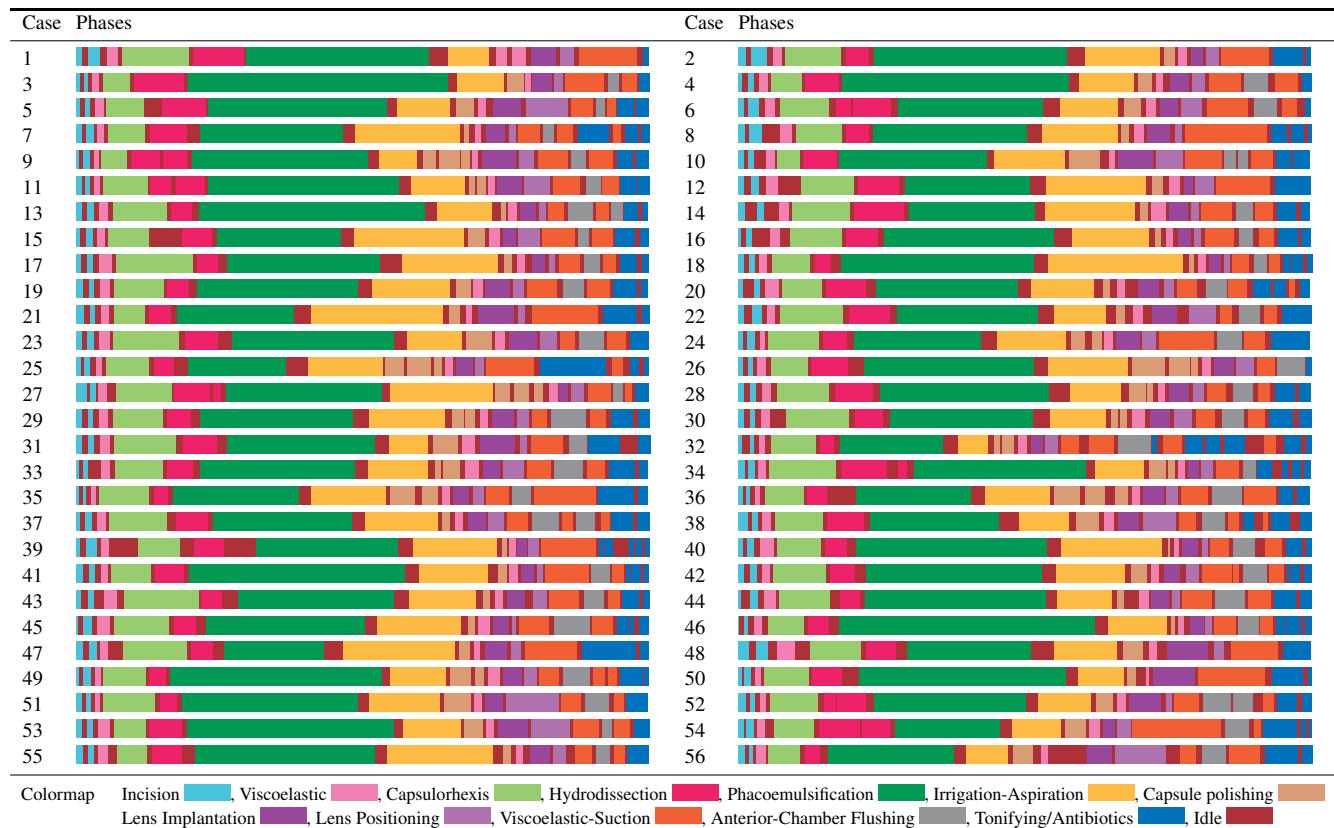


Figure 2. Total duration of the annotated phases in the 56 annotated cataract surgery videos (in seconds).

- As shown in Figure 2, phase recognition datasets for cataract surgery are extremely imbalanced, as the longest phase (phacoemulsification) and the shortest phase (incision) cover 28.72% and 2.1% of the annotations, respectively.
- Videos may exhibit defocus blur stemming from manual camera focus adjustments²³.
- Unintentional eye movements and rapid instrument motions close to the camera result in motion blur, impairing distinctive spatial details.
- Lack of metadata in stored videos precludes additional contextual information.
- Variances in patients' eye visuals generate substantial inter-video distribution disparities, demanding ample training data

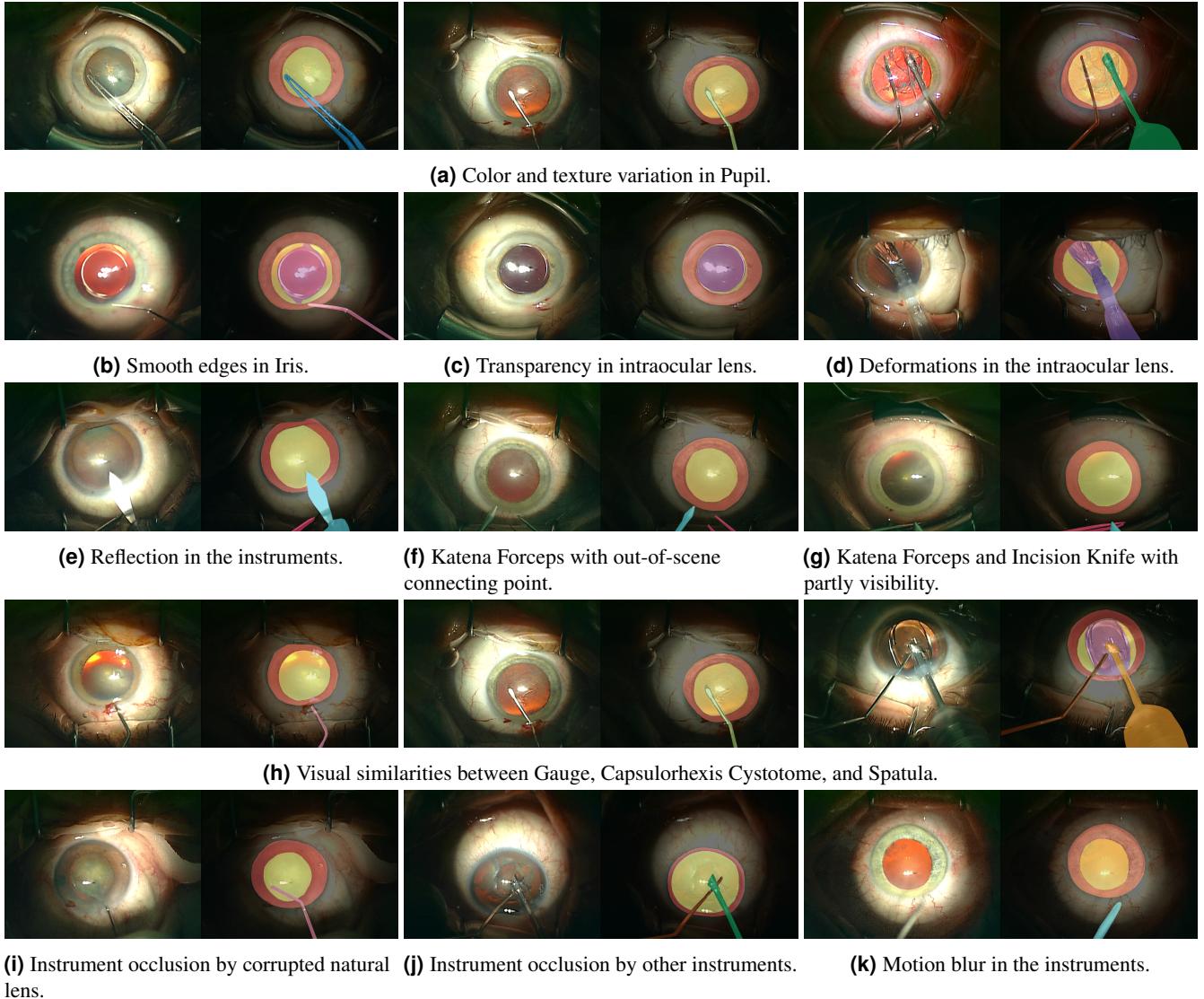


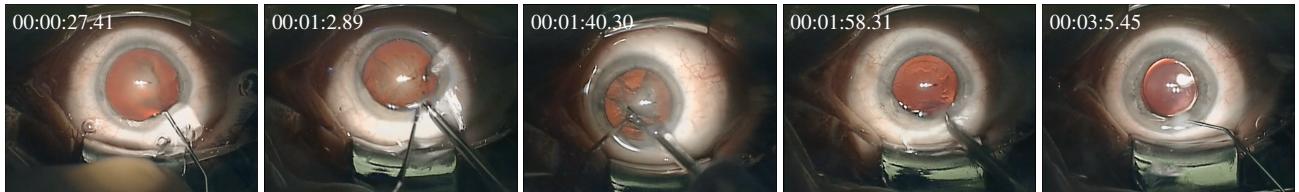
Figure 3. Visualization of pixel-based annotations corresponding to relevant anatomical structures and instruments in cataract surgery and the challenges associated with different object (Iris: , Pupil: , Lens: , Slit/Incision Knife: , Gauge: , Spatula: , Capsulorhexis Cystotome: , Phacoemulsifier Tip: , Irrigation-Aspiration: , Lens Injector : , Capsulorhexis Forceps: , Katena Forceps:).

to build networks with generalizable performance.

As shown in Figure 1, regular cataract surgery can include twelve action phases, including incision, viscoelastic, capsulorhexis, hydrodissection, phacoemulsification, irrigation-aspiration, capsule polishing, lens implantation, lens positioning, viscoelastic-suction, anterior-chamber flushing, and tonifying/antibiotics. Besides, the idle phases refer to the time spans in the middle of a phase or between two phases when the surgeons mainly change the instruments and no instrument is visible inside the frames. We provide a large annotated dataset to enable comprehensive studies on deep-learning-based phase recognition in cataract surgery videos.

Table 1 visualizes the phase annotations corresponding to 56 regular cataract surgery videos, with a spatial resolution of 1024×768 , a temporal resolution of 30 fps, an average duration of 6.45 minutes, and a standard deviation of 2.04. This dataset comprises patients with an average age of 75 years, ranging from 51 to 93 years, and a standard deviation of 8.69 years. The videos present in the phase recognition dataset correspond to surgeries executed by surgeons with an average experience of 8929 surgeries and a standard deviation of 6350 surgeries.

Semantic segmentation dataset. Figure 3 visualizes pixel-level annotations for relevant anatomical objects and instruments. As illustrated in Figure 3, semantic segmentation in cataract surgery videos poses the following challenges^{21,22,24}:



(a) Pupil contraction during cataract surgery.



(b) Severe clockwise IOL rotations.

Figure 4. Intra-operative irregularities in cataract surgery.

- Variations in color, shape, size, and texture in pupil,
- Transparency and deformations in the artificial lens,
- Smooth edges and color variations in iris,
- Occlusion, motion blur, reflection, and partly visibility in instruments,
- Visual similarities between different instruments in case of multi-class instrument segmentation,

The semantic segmentation dataset includes frames from 30 regular cataract surgery videos with a spatial resolution of 1024×768 . Frame extraction is performed at the rate of one frame per five seconds. Subsequently, the frames featuring very harsh motion blur or out-of-scene iris are excluded from the dataset. We provide pixel-level annotations for three relevant anatomical structures, including the iris, pupil, and intraocular lens, as well as nine instruments used in regular cataract surgeries, including slit/incision knife, gauge, spatula, capsulorhexis cystome, phacoemulsifier tip, irrigation-aspiration, lens injector, capsulorhexis forceps, and katana forceps. All annotations are performed using polygons in the Supervisely platform³, and exported as JSON files. Within this dataset, the included individuals possess an average age of 74.5 years, spanning from 51 to 90 years, with a standard deviation of 8.43 years. Additionally, the videos contained in the phase recognition dataset depict surgeries conducted by surgeons whose collective experience averages 8033 surgeries, with a standard deviation of 3894 surgeries. The provided dataset enables a reliable study of segmentation performance for relevant anatomical structures, binary instruments, and multi-class instruments.

Irregularity detection dataset. This dataset contains two small subsets of major intra-operative irregularities in cataract surgery, including pupil reaction¹⁷ and lens rotation¹⁶.

• *Pupil Contraction:* During the phacoemulsification phase, where the occluded natural lens is fragmented and suctioned, there exists a heightened risk of causing damage to the delicate iris. Even very subtle trauma to the tissue can lead to undesirable pupil constriction²⁵. These sudden reactions in pupil size can lead to serious intra-operative implications. Especially during the phacoemulsification phase, where the instrument is deeply inserted inside the eye, sudden changes in pupil size may lead to injuries to the eye's tender tissues. Besides, achieving precise IOL alignment or centration becomes challenging in cases where intraoperative pupil contraction (miosis) occurs. Particularly in multifocal IOLs, minor displacements or tilts, which might be negligible for conventional mono-focal IOLs, can significantly compromise visual performance. In the case of toric IOLs, precise alignment of the torus is crucial, as any deviation diminishes the IOL's effectiveness. Detection of unusual pupil reactions and severe pupil contractions during the surgery can highly contribute to the overall outcomes of cataract surgery and provide important insight for further post-operative investigations. Figure 4-top demonstrates an example of severe pupil contraction during cataract surgery.

• *IOL Rotation:* Although aligned and centered upon surgery's conclusion, the IOL may rotate or dislocate following the surgery. Even slight deviations, such as minor misalignments of the torus in toric IOLs or the slight displacement and tilting of multifocal IOLs, can result in significant distortions in vision and leave patients dissatisfied. The sole way to

³<https://supervisely.com/>

address this postoperative complication is follow-up surgery, which entails added costs, heightened surgical risks, and patient discomfort. Identification of intra-operative indicators for predicting and preventing post-surgical IOL dislocation is an unmet clinical need. It is argued that intra-operative rotation of IOLs during cataract surgery is the leading cause of post-operative misalignments²⁶. Hence, automatic detection and measurement of intra-operative lens rotations can effectively contribute to preventing post-operative IOL dislocation. Figure 4-bottom represents fast clockwise rotations of IOL during unfolding, which occur in less than seven seconds.

Experimental Settings for Phase Recognition

Network Architectures. We adopt a combined CNN-RNN framework for phase recognition. The CNN component, serving as the backbone model, is responsible for the extraction of distinctive features from individual frames within the video sequence. To achieve this, two different pre-trained CNN architectures, VGG16 and ResNet50, are employed. The output feature map of the CNN is fed into a recurrent neural network (RNN). The RNN component focuses on capturing temporal features from the input video clip. We compare the performance of four different RNN architectures, including LSTM, GRU, BiLSTM, and BiGRU.

Training Settings. We adopt a one-versus-rest strategy to evaluate phase recognition performance^{14,27}. Accordingly, we segment all videos corresponding to each phase into three-second clips with an overlap of one second. Afterward, the entire dataset is split into two categories: the designated target phase and the remaining phases (the "rest" class). We apply offline augmentations to the videos across all categories. Typically, the number of clips in the target category is significantly lower than in the rest category. To rectify this imbalance problem, we employ a random selection process from the "rest" category, aligning it with the clip count in the target category. This strategy ensures an equivalent number of clips in both classes. The employed augmentations include gamma and contrast adjustments with a factor of 0.5, Gaussian blur with a sigma of 10, random rotation up to 20 degrees, brightness within a range of $[-0.3, 0.3]$, and saturation within a range of $[0.5, 1.5]$. To maximize diversity within our training set, we employ a random sampling strategy during training. Specifically, we configure the network's input sequence to comprise 10 frames randomly selected from 90 frames within each three-second clip. In all settings, the backbone network employed for feature extraction is pre-trained on the ImageNet dataset. The RNN component is constructed with a single recurrent layer comprising 64 units. This is followed by a dense layer with 64 units, and finally, a two-unit layer with a Softmax activation function. To mitigate the risk of overfitting, the last four layers of the CNN component are kept frozen during training, and dropout regularization with a rate of 0.5 is applied to the output feature map of the recurrent layer. All models are trained on 32 videos and tested on non-overlapping clips from the remaining videos. We use a binary cross-entropy loss function and Adam optimizer, a learning rate equal to 0.001, and a batch size of 16. The network's input dimensions are set to 224×224 . We compare the performance of the trained models using accuracy and F1 score.

Experimental Settings for Semantic Segmentation

Network Architectures. We perform experiments to validate the robustness of our pixel-level annotations using several state-of-the-art baselines targeting general images, medical images, and surgical videos. The specifications of the baselines are listed in Table 2.

Training Settings. For all neural networks, the backbones are initialized with ImageNet's pre-trained parameters³⁶. We train all networks with a batch size of eight and set the initial learning rate to 0.001, which decreases during training using polynomial decay $lr = lr_{init} \times (1 - \frac{iter}{total_iter})^{0.9}$. The input size of the networks is set to 512×512 . We apply cropping and

Table 2. Specifications of the proposed and alternative approaches. In "Upsampling" column, "Trans Conv" stands for *Transposed Convolution*.

Model	Backbone	Params.	Upsampling	Target	Reference
DeepPyramid	VGG16	33.57 M	Bilinear	Medical Images	²¹
Adapt-Net	VGG16	24.69 M	Bilinear	Medical Images	¹⁶
UNet++	VGG16	24.24 M	Bilinear	Medical Images	²⁸
ReCal-Net	VGG16	22.93 M	Bilinear	Medical Images	²²
CPFNet	VGG16 ResNet34	39.17 M 34.66 M	Bilinear	Medical Images	²⁹
CE-Net	VGG16 ResNet34	33.50 M 29.90 M	Trans Conv	Medical Images	³⁰
FED-Net	ResNet50	59.52 M	Trans Conv & PixelShuffle	Liver Lesion	³¹
scSENet	VGG16 ResNet34	22.90 M 25.25 M	Bilinear	Medical Images	³²
DeepLabV3+	ResNet50	26.68 M	Bilinear	Scene	³³
UPerNet	ResNet50	51.26 M	Bilinear	Scene	³⁴
U-Net+ ⁴	VGG16	22.55 M	Bilinear	Medical Images	³⁵

Table 3. Number of instances and presence in the frames (% of total number of frames in each fold).

Category	Class Name	All Videos	Fold1	Fold2	Fold3	Fold4	Fold5
Anatomy	Iris	2256 (100.0%)	561 (100.0%)	459 (100.0%)	420 (100.0%)	385 (100.0%)	431 (100.0%)
	Pupil	2256 (100.0%)	561 (100.0%)	459 (100.0%)	420 (100.0%)	385 (100.0%)	431 (100.0%)
	Intraocular Lens	537 (23.8%)	107 (19.07%)	119 (25.93%)	102 (24.29%)	106 (27.53%)	103 (23.9%)
Instruments	Slit/Incision Knife	50 (2.22%)	12 (2.14%)	10 (2.18%)	12 (2.86%)	4 (1.04%)	12 (2.78%)
	Gauge	426 (18.88%)	103 (18.36%)	90 (19.61%)	79 (18.81%)	76 (19.74%)	78 (18.1%)
	Spatula	728 (32.27%)	214 (38.15%)	132 (28.76%)	148 (35.24%)	105 (27.27%)	129 (29.93%)
	Capsulorhexis Cystotome	85 (3.77%)	20 (3.57%)	18 (3.92%)	12 (2.86%)	11 (2.86%)	24 (5.57%)
	Phacoemulsifier Tip	547 (24.25%)	148 (26.38%)	91 (19.83%)	101 (24.05%)	95 (24.68%)	112 (25.99%)
	Irrigation-Aspiration	456 (20.21%)	122 (21.75%)	91 (19.83%)	98 (23.33%)	71 (18.44%)	74 (17.17%)
	Lens Injector	66 (2.93%)	14 (2.5%)	11 (2.4%)	14 (3.33%)	13 (3.38%)	14 (3.25%)
	Capsulorhexis Forceps	108 (4.79%)	33 (5.88%)	21 (4.58%)	22 (5.24%)	21 (5.45%)	11 (2.55%)
	Katena Forceps	29 (1.29%)	8 (1.43%)	3 (0.65%)	8 (1.9%)	3 (0.78%)	7 (1.62%)
	All	1778 (78.81%)	462 (82.35%)	345 (75.16%)	344 (81.9%)	296 (76.88%)	331 (76.8%)

Table 4. Average pixels corresponding to different labels per frame.

Category	Class Name	All Videos	Fold1	Fold2	Fold3	Fold4	Fold5
Anatomy	Iris	45939	41874	47792	44867	47963	48494
	Pupil	36013	38594	33578	35900	35291	35999
	Intraocular Lens	9135	7056	10017	9153	10405	9748
Instruments	Slit/Incision Knife	1140	1088	1179	1163	1206	1086
	Gauge	299	222	337	454	168	326
	Spatula	2613	3163	2078	2893	2309	2466
	Capsulorhexis Cystotome	5523	4760	4773	6551	6345	5580
	Phacoemulsifier Tip	5230	4388	5526	7646	4451	4354
	Irrigation-Aspiration	1083	790	1138	1311	1153	1123
	Lens Injector	512	465	543	673	318	556
	Capsulorhexis Forceps	172	288	104	225	23	176
	Katena Forceps	823	906	678	1065	1133	357
	All	17397	16069	16357	21981	17105	16025

random rotation (up to 30 degrees), color jittering (brightness = 0.7, contrast = 0.7, saturation = 0.7), Gaussian blurring, and random sharpening as augmentations during training, and use the *cross entropy log dice* loss during training as in eq. (1),

$$\mathcal{L} = (\lambda) \times BCE(\mathcal{X}_{true}(i, j), \mathcal{X}_{pred}(i, j)) - (1 - \lambda) \times (\log \frac{2 \sum \mathcal{X}_{true} \odot \mathcal{X}_{pred} + \sigma}{\sum \mathcal{X}_{true} + \sum \mathcal{X}_{pred} + \sigma}), \quad (1)$$

where \mathcal{X}_{true} denotes the ground truth binary mask, and \mathcal{X}_{pred} denotes the predicted mask ($0 \leq \mathcal{X}_{pred}(i, j) \leq 1$). The parameter $\lambda \in [0, 1]$ is set to 0.8 in our experiments, and \odot refers to the Hadamard product (element-wise multiplication). Besides, the parameter σ is the Laplacian smoothing factor, which is added to (i) prevent division by zero and (ii) avoid overfitting (in experiments, $\sigma = 1$). We compare the performance of baselines using average dice and average intersection over union (IoU).

Data Records

All datasets and annotations will be publicly released in Synapse upon the acceptance of the paper (accessible for anonymous review in Figshare).

Frame-level annotations for phase recognition are provided in CSV files, determining the first and the last frames for all action phases per video. The preprocessing codes to extract all action and idle phases from a video using the CSV files are provided in the GitHub repository of the paper. Table 1 visualizes our phase annotations for 56 cataract surgery videos. Furthermore, Figure 2 demonstrates the total duration of the annotations corresponding to each phase from 56 videos.

Pixel-level annotations are provided in two formats: (1) Supervisely format, for which we provide Python codes for mask creation from JSON files, and (2) COCO format, which also provides bounding box annotations for all pixel-level annotated

objects. The latter annotations can be used for object localization problems. The preprocessing codes to create training masks for "anatomy plus instrument segmentation", "binary instrument segmentation", and "multi-class instrument segmentation" are provided in the GitHub repository of the paper. We have formed five folds with patient-wise separation, meaning every fold consists of the frames corresponding to six distinct videos. Table 3 compares the number of instances and their appearance percentage in the frames. Besides, Table 4 lists the average number of pixels per frame corresponding to each label.

Technical Validation

Table 5 showcases the phase recognition performance of several CNN-RNN architectures. In our evaluations, we have combined the phases of viscoelastic and anterior-chamber flushing due to their shared visual features. The collective findings reveal commendable and satisfactory phase recognition performance across diverse backbones and recurrent network setups. Notably, the incorporation of bidirectional recurrent layers has consistently amplified detection accuracy and F1-Score across all configurations.

Table 5. Phase recognition performance of several CNN-RNN architectures.

Network	ResNet50-LSTM		ResNet50-GRU		ResNet50-BiLSTM		ResNet50-BiGRU	
Phases	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)
Incision	83.35	81.48	83.33	83.31	86.67	86.61	86.67	86.61
Viscoelastic/AC Flushing	65.12	64.81	69.77	69.17	62.21	59.43	60.47	57.42
Capsulorhexis	85.71	85.54	86.05	85.85	90.14	90.09	90.82	90.81
Hydrodissection	88.27	88.22	86.42	86.16	88.89	88.81	87.04	86.86
Phacoemulsification	95.17	95.17	94.16	94.16	95.17	95.37	94.67	94.67
Irrigation-Aspiration	89.91	89.88	87.39	87.32	86.47	86.44	87.84	87.82
Capsule Polishing	86.17	85.90	81.91	81.66	88.30	88.30	87.23	87.21
Lens Implantation	85.14	84.80	81.08	80.38	86.49	86.24	90.54	90.50
Lens Positioning	87.50	87.49	87.50	87.45	92.19	92.14	89.06	89.00
Viscoelastic-Suction	91.73	91.72	89.10	89.03	90.23	90.15	90.98	90.97
Tonifying/Antibiotics	85.83	85.75	81.67	81.33	86.67	86.66	88.33	88.30
Network	VGG16-LSTM		VGG16-GRU		VGG16-BiLSTM		VGG16-BiGRU	
Phases	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)	Accuracy (%)	F1-Score (%)
Incision	83.33	82.86	86.67	86.43	86.67	86.43	90.00	89.90
Viscoelastic/AC Flushing	64.53	63.18	63.37	62.30	64.53	63.89	66.28	64.91
Capsulorhexis	87.07	87.04	87.76	87.73	88.44	88.42	89.80	89.79
Hydrodissection	86.42	86.42	85.23	85.79	87.89	88.89	87.04	87.02
Phacoemulsification	93.86	93.86	93.36	93.36	93.26	93.26	92.86	92.86
Irrigation-Aspiration	86.70	86.55	86.47	86.31	88.53	88.51	88.53	88.48
Capsule Polishing	87.23	87.23	86.17	85.90	90.43	90.37	88.30	88.14
Lens Implantation	82.43	82.04	86.49	86.24	83.78	83.35	85.14	85.00
Lens Positioning	85.94	85.93	82.81	82.47	87.50	87.49	84.38	84.13
Viscoelastic-Suction	81.95	81.72	78.95	78.24	82.71	82.31	82.33	81.90
Tonifying/Antibiotics	82.50	82.50	83.33	83.33	81.67	81.33	85.00	84.96

Table 6. Quantitative evaluations of "anatomy plus instrument" segmentation performance for neural network architectures listed in Table 2.

Object	Backbone	Network	Iris		Pupil		Lens		Instruments	
			IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)
VGG16	CPNNet	UNet+	85.43	92.13	89.50	94.46	79.39	88.47	63.37	77.56
		scSENet	85.52	92.18	89.34	94.37	78.89	88.18	63.53	77.68
		FEDNet	84.09	91.34	88.25	93.76	77.93	87.56	60.12	75.07
		CE-Net	82.98	90.68	86.54	92.78	73.40	84.53	56.50	72.17
	AdaptNet	85.10	91.93	89.50	94.45	80.66	89.26	62.90	77.21	
		UNetPP	85.20	91.99	89.46	94.43	79.07	88.26	63.63	77.76
		AdaptNet	85.50	92.18	90.29	94.89	83.02	90.70	62.18	76.68
		ReCal-Net	85.33	92.07	90.19	94.84	82.53	90.43	62.55	76.95
		DeepPyramid	86.10	92.52	90.61	95.07	83.72	91.11	63.91	77.98
ResNet34	CPNNet	scSENet	84.77	91.74	87.72	93.42	65.31	76.02	27.11	34.64
		FEDNet	81.83	89.99	84.13	91.35	51.75	65.61	53.51	69.69
		CE-Net	76.52	86.43	84.84	91.80	66.21	79.48	30.88	43.86
		CPFNet	83.59	91.05	88.74	94.03	81.24	89.63	61.71	76.31
ResNet50	UPerNet	UPerNet	85.15	91.97	90.03	94.75	83.72	91.11	63.48	77.65
		DeepLabV3+	79.97	88.83	86.24	92.61	77.23	87.12	53.53	69.67

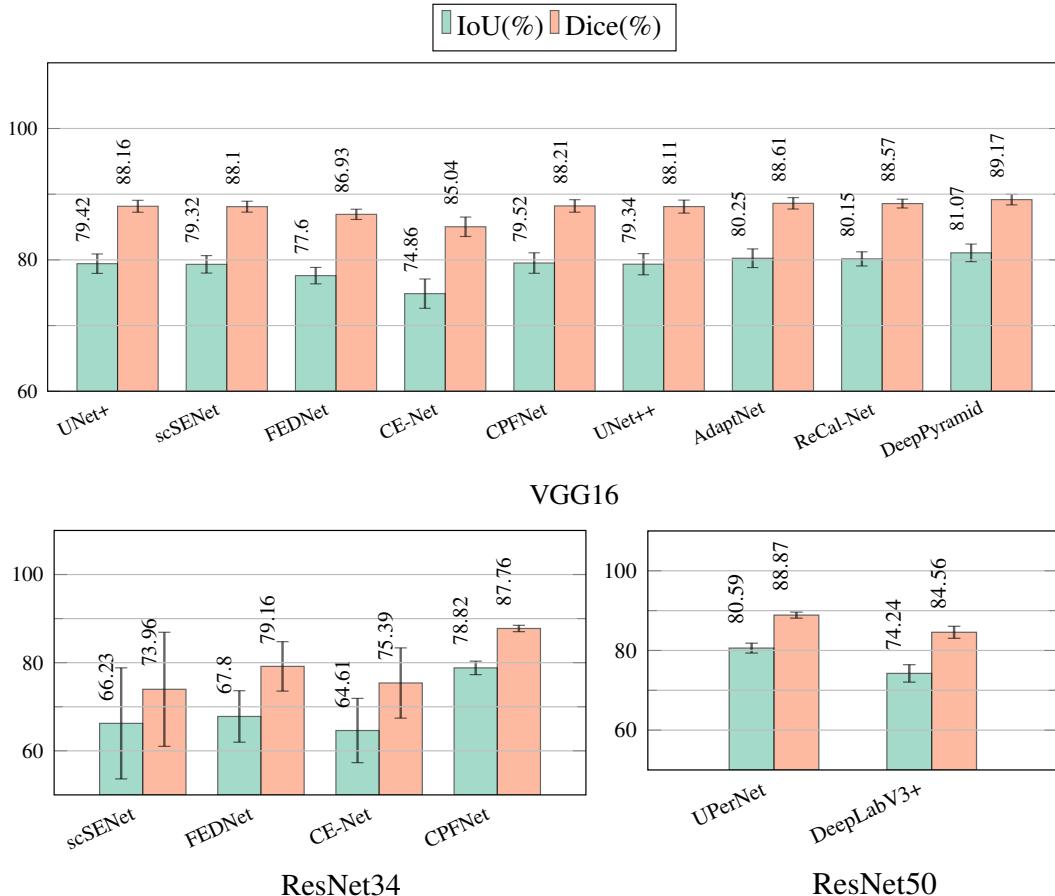


Figure 5. Average and standard deviation of "anatomy plus instrument" segmentation results for neural network architectures listed in Table 2.

Furthermore, networks leveraging the ResNet50 backbone display marginally superior performance compared to those utilizing VGG16. This outcome can be attributed to the deeper architecture of ResNet50, facilitating the extraction of intricate features essential for accurate recognition. The results also reveal the distinguishability of different phases in cataract surgery. Precisely, the phacoemulsification phase consistently attains the highest accuracy and F1 score, attributed to the distinctive phacoemulsification instrument and the unique texture of the pupil during this phase. Conversely, the least robust detection performance aligns with the viscoelastic/AC flushing phases, accentuating the visual resemblances shared between these phases and other phases within cataract surgery videos.

Table 6 provides a quantitative analysis of "anatomy plus instrument" segmentation performance for various neural network architectures. The results notably highlight that segmenting the relevant anatomical structures emerges as a comparatively less challenging task than instrument segmentation for all networks. Specifically, the best performance corresponds to pupil segmentation, attributable to its distinct features and sharp boundaries. In contrast, lens segmentation demonstrates relatively lower performance due to its transparent nature and an inherent imbalance issue (outlined in Table 3). The segments involving instruments, however, confront significant challenges. This class is marked by major distortions, encompassing motion blur, reflections, and occlusions, collectively contributing to the relatively low performance of the networks. The best performance corresponds to the DeepPyramid network with a VGG16 backbone, consistently yielding optimal results across all classes.

Figure 5 visually compares the Dice and IoU metrics' averages and standard deviations across five folds for the evaluated neural networks. According to the results, DeepPyramid, AdaptNet, and ReCal-Net are the three best-performing networks for anatomy and instrument segmentation in cataract surgery videos.

Within Table 7, a thorough comparison is made between the performance of various neural network architectures concerning intra-domain and cross-domain scenarios. These architectures are trained using our binary instrument annotations. The results clearly indicate statistical differences between Cataract-1k and CaDIS datasets. Concretely, the average dice coefficient for binary instrument segmentation equals 77% within the Cataract-1k dataset. However, this performance metric markedly contracts, remaining confined to around 67% (with AdaptNet illustrating 66.23%) when extended to the CaDIS dataset. This

Table 7. Single domain and cross-domain binary instrument segmentation performance for neural network architectures listed in Table 2.

Domain		Source (Cataract-1K)		Target (CaDIS)	
Backbone	Network	IoU (%)	Dice (%)	IoU (%)	Dice (%)
VGG16	UNet+	71.58 \pm 3.03	79.06 \pm 2.53	29.89 \pm 2.48	39.19 \pm 2.84
	scSENet	71.53 \pm 3.57	79.05 \pm 2.99	25.56 \pm 1.97	34.11 \pm 2.32
	FEDNet	69.40 \pm 3.68	77.62 \pm 3.07	21.45 \pm 4.35	29.08 \pm 5.54
	CE-Net	71.27 \pm 4.57	79.65 \pm 3.98	14.46 \pm 5.16	20.17 \pm 6.40
	CPFNet	78.38 \pm 2.53	86.00 \pm 1.99	16.86 \pm 5.42	22.84 \pm 6.64
	UNetPP	71.66 \pm 3.48	79.15 \pm 2.93	30.54 \pm 1.50	40.01 \pm 1.65
	AdaptNet	74.42 \pm 3.25	81.49 \pm 2.64	49.55 \pm 1.49	61.65 \pm 1.63
	ReCal-Net	66.88 \pm 5.66	74.15 \pm 5.67	37.99 \pm 4.15	49.02 \pm 4.64
	DeepPyramid	77.97 \pm 3.78	84.95 \pm 3.02	49.43 \pm 2.06	60.79 \pm 1.96
ResNet34	scSENet	77.30 \pm 2.79	84.64 \pm 2.21	44.36 \pm 1.51	55.26 \pm 1.62
	FEDNet	76.86 \pm 2.66	85.01 \pm 2.05	40.46 \pm 1.35	51.32 \pm 1.42
	CE-Net	34.78 \pm 2.88	47.29 \pm 2.90	36.55 \pm 2.59	50.61 \pm 2.82
	CPFNet	44.92 \pm 1.20	56.17 \pm 1.50	43.71 \pm 1.93	57.43 \pm 2.01
	AdaptNet	75.10 \pm 2.99	82.34 \pm 2.43	54.15 \pm 0.80	66.23 \pm 0.80
	ReCal-Net	69.76 \pm 5.96	77.27 \pm 5.98	48.66 \pm 2.35	60.36 \pm 2.84
ResNet50	UPerNet	78.36 \pm 2.72	85.51 \pm 2.13	40.28 \pm 1.33	50.82 \pm 1.52
	DeepLabV3+	68.77 \pm 2.31	78.14 \pm 2.13	30.72 \pm 4.95	41.50 \pm 5.94

considerable variance starkly underscores the substantial domain shift inherently present between these two datasets. These results demonstrate the necessity of strategically exploring semi-supervised and domain adaptation techniques to elevate instrument segmentation performance in cataract surgery videos with cross-dataset domain shifts³⁷.

Usage Notes

The datasets are licensed under CC BY. For further legal details, we kindly request the readers to refer to the complete [license terms](#).

Besides, anyone can view the sample videos and images from the dataset and access the GitHub repository for dataset preparation codes.

Code availability

We provide all code for mask creation using JSON annotations and phase extraction using CSV files, as well as their usage instruction in the GitHub repository of this paper (accessible for anonymous review in Figshare).

Acknowledgements

We would like to thank Daniela Stefanics for helping us in annotating the datasets to the highest quality.

This work was funded by Haag-Streit Foundation Switzerland and the FWF Austrian Science Fund under grant P 32010-N38. This work is performed under ethics committee approval (EK 28/17).

Author contributions statement

N.G. wrote the original draft. R.S., M.Z., S.W., K.S., Y.E., and D.P. acquired the projects' funding. R.S. and K.S. were responsible for the projects' supervision. N.G. and K.S. organized the annotation process. Y.E. and D.P. provided expert information on the cataract surgery phases, relevant anatomical structures, and instruments used in regular cataract surgery videos. N.G. and D.P. reviewed and corrected the annotations. N.G. designed, implemented, and evaluated semantic segmentation experiments of the technical validation. N.G. and S.N. designed phase recognition experiments of technical validation. S.N. implemented and evaluated phase recognition experiments of technical validation. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Ma, L. *et al.* Simulation of postoperative facial appearances via geometric deep learning for efficient orthognathic surgical planning. *IEEE Transactions on Med. Imaging* **42**, 336–345, [10.1109/TMI.2022.3180078](https://doi.org/10.1109/TMI.2022.3180078) (2023).
2. Quon, J. *et al.* Deep learning for automated delineation of pediatric cerebral arteries on pre-operative brain magnetic resonance imaging. *front surg* 2020; 7 (2020).
3. Xiao, D. *et al.* Estimating reference bony shape models for orthognathic surgical planning using 3d point-cloud deep learning. *IEEE J. Biomed. Heal. Informatics* **25**, 2958–2966, [10.1109/JBHI.2021.3054494](https://doi.org/10.1109/JBHI.2021.3054494) (2021).
4. Yanik, E. *et al.* Deep neural networks for the assessment of surgical skills: A systematic review. *The J. Def. Model. Simul.* **19**, 159–171 (2022).
5. Lam, K. *et al.* Machine learning for technical skill assessment in surgery: a systematic review. *NPJ digital medicine* **5**, 24 (2022).
6. Wang, Z. & Majewicz Fey, A. Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *Int. journal computer assisted radiology surgery* **13**, 1959–1970 (2018).
7. Wang, Z. & Fey, A. M. Satr-dl: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1793–1796 (IEEE, 2018).
8. Soleymani, A. *et al.* Surgical skill evaluation from robot-assisted surgery recordings. In *2021 International Symposium on Medical Robotics (ISMR)*, 1–6 (IEEE, 2021).
9. Aksamentov, I., Twinanda, A. P., Mutter, D., Marescaux, J. & Padoy, N. Deep neural networks predict remaining surgery duration from cholecystectomy videos. In *Medical Image Computing and Computer-Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part II 20*, 586–593 (Springer, 2017).
10. Twinanda, A. P., Yenger, G., Mutter, D., Marescaux, J. & Padoy, N. Rsdnet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations. *IEEE transactions on medical imaging* **38**, 1069–1078 (2018).
11. Marafioti, A. *et al.* Catanet: predicting remaining cataract surgery duration. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, 426–435 (Springer, 2021).
12. Ghamsarian, N. Enabling relevance-based exploration of cataract videos. In *Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR '20*, 378–382, [10.1145/3372278.3391937](https://doi.org/10.1145/3372278.3391937) (2020).
13. Burton, M. J. *et al.* The lancet global health commission on global eye health: vision beyond 2020. *The Lancet Glob. Heal.* **9**, e489–e551 (2021).
14. Ghamsarian, N., Taschwer, M., Putzgruber-Adamitsch, D., Sarny, S. & Schöffmann, K. Relevance detection in cataract surgery videos by spatio-temporal action localization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, 10720–10727 (2021).
15. Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M. & Schöffmann, K. Relevance-based compression of cataract surgery videos using convolutional neural networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3577–3585 (2020).
16. Ghamsarian, N. *et al.* Lensid: A cnn-rnn-based framework towards lens irregularity detection in cataract surgery videos. In de Bruijne, M. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 76–86 (Springer International Publishing, Cham, 2021).
17. Sokolova, N. *et al.* Automatic detection of pupil reactions in cataract surgery videos. *Plos one* **16**, e0258390 (2021).
18. Ghamsarian, N. *et al.* Predicting postoperative intraocular lens dislocation in cataract surgery via deep learning. *arXiv preprint arXiv:2312.03401* (2023).
19. Al Hajj, H. *et al.* Cataracts: Challenge on automatic tool annotation for cataract surgery. *Med. image analysis* **52**, 24–41 (2019).
20. Grammatikopoulou, M. *et al.* Cadis: Cataract dataset for surgical rgb-image segmentation. *Med. Image Analysis* **71**, 102053 (2021).

21. Ghamsarian, N., Taschwer, M., Sznitman, R. & Schoeffmann, K. Deep pyramid: Enabling pyramid view and deformable pyramid reception for semantic segmentation in cataract surgery videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 276–286 (Springer, 2022).
22. Ghamsarian, N. *et al.* Recal-net: Joint region-channel-wise calibrated network for semantic segmentation in cataract surgery videos. In Mantoro, T., Lee, M., Ayu, M. A., Wong, K. W. & Hidayanto, A. N. (eds.) *Neural Information Processing*, 391–402 (Springer International Publishing, Cham, 2021).
23. Ghamsarian, N., Taschwer, M. & Schoeffmann, K. Deblurring cataract surgery videos using a multi-scale deconvolutional neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 872–876 (2020).
24. Ghamsarian, N., Wolf, S., Zinkernagel, M., Schoeffmann, K. & Sznitman, R. Deep pyramid+: Medical image segmentation using pyramid view fusion and deformable pyramid reception. *arXiv preprint arXiv:2312.03409* (2023).
25. Mirza, S. A., Alexandridou, A., Marshall, T. & Stavrou, P. Surgically induced miosis during phacoemulsification in patients with diabetes mellitus. *Eye* **17**, 194–199, [10.1038/sj.eye.6700268](https://doi.org/10.1038/sj.eye.6700268) (2003).
26. Oshika, T. *et al.* Prospective assessment of plate-haptic rotationally asymmetric multifocal toric intraocular lens with near addition of + 1.5 diopters. *BMC Ophthalmol.* **20**, 454, [10.1186/s12886-020-01731-3](https://doi.org/10.1186/s12886-020-01731-3) (2020).
27. Nasirihaghghi, S., Ghamsarian, N., Stefanics, D., Schoeffmann, K. & Husslein, H. Action recognition in video recordings from gynecologic laparoscopy. In *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, 29–34, [10.1109/CBMS58004.2023.00187](https://doi.org/10.1109/CBMS58004.2023.00187) (2023).
28. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Med. Imaging* **39**, 1856–1867, [10.1109/TMI.2019.2959609](https://doi.org/10.1109/TMI.2019.2959609) (2020).
29. Feng, S. *et al.* Cpfnet: Context pyramid fusion network for medical image segmentation. *IEEE Transactions on Med. Imaging* **39**, 3008–3018, [10.1109/TMI.2020.2983721](https://doi.org/10.1109/TMI.2020.2983721) (2020).
30. Gu, Z. *et al.* Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Med. Imaging* **38**, 2281–2292, [10.1109/TMI.2019.2903562](https://doi.org/10.1109/TMI.2019.2903562) (2019).
31. Chen, X., Zhang, R. & Yan, P. Feature fusion encoder decoder network for automatic liver lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 430–433, [10.1109/ISBI.2019.8759555](https://doi.org/10.1109/ISBI.2019.8759555) (2019).
32. Roy, A. G., Navab, N. & Wachinger, C. Recalibrating fully convolutional networks with spatial and channel “squeeze and excitation” blocks. *IEEE Transactions on Med. Imaging* **38**, 540–549 (2019).
33. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818 (2018).
34. Xiao, T., Liu, Y., Zhou, B., Jiang, Y. & Sun, J. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, 418–434 (2018).
35. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 234–241 (2015).
36. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).
37. Ghamsarian, N. *et al.* Domain adaptation for medical image segmentation using transformation-invariant self-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 331–341 (Springer, 2023).