Real-Time Intelligence using Deep Learning for Skill Enhancement in Capsulorhexis Cataract Surgery

Geethika Golamari, Aditya Sharma, Brinda A, C V M Yeshaswini, Manjunath K P, B Niranjana Krupa
PES University, Bengaluru, India
{geethikagolamari, adiabhi1110, brinda.abhilash, yeshascvm408, manjunathkpgupta}@gmail.com, bnkrupa@pes.edu,

***Abstract*—This research focuses on creating a guidance map for computer-assisted cataract surgeries. The main focus is to improve surgery duration and assist young practitioners while performing cataract surgeries with the capsulorhexis method. The research features two phases, namely, the Incision Phase and the Capsulorhexis Phase among thirteen different phases of cataract surgery. This research provides methods for real-time phase recognition and segmentation. The study uses ResNet-18 for phase recognition to identify the Incision Phase and the Capsulorhexis Phase. The model was able to achieve an accuracy of 94.42% and F1 Score of 0.9452. UNet with Guidance Filter Module (GFM) is used to segment the cornea and the main incision knife. The UNet model produced masks with a mean IoU of 0.9315, mean Dice of 0.9642 with an inference time of 6.2ms for incision knife segmentation. For the task of cornea segmentation, the model produced results with an inference time of 6.5ms with mean IoU score of 0.9450 and mean Dice score of 0.9716. Using the extracted features and the recognized surgical phase, a guidance map is plotted for each frame of the video. These frames are rendered at 30 Frames Per Second (FPS) to create a video. The Incision Phase map provides a guide for the main incision and the side ports along with horizontal depth tracking of the main incision. The capsulorhexis guidance map provides a guide along the rhexis circle (circle along which capsule is torn).**

***Index Terms*—Cataract Surgery, Capsulorhexis, Incision, Depth Tracking, ResNet, UNet, Phase Recognition, Segmentation.**

I. Introduction

One of the leading causes of blindness in the world is cataract. A cataract occurs when the normally transparent lens within the eye develops a cloudy appearance due to the gradual breakdown of proteins in the lens. This cloudiness causes vision to be blurry, hazy, or less vivid eventually leading to blindness. At least 26 million cataract surgeries are performed yearly [1]. Major logistical and organizational issues are faced due to this high demand sparking a need for automatic systems to enable faster decision-making for less experienced surgeons. Thus, the development of artificial intelligence-based solutions began to fulfil this requirement to analyze and interpret surgical videos. Based on this evident requirement, the problem statement for this study can be formulated as: 'Integration of surgical phase recognition, horizontal depth tracking of the incision and segmentation to create a guidance map for improved outcomes and continuous skill enhancement in cataract surgery'.

To fulfil this problem statement, the surgical phase, anatomical structures, and instruments are first identified in a surgical scene. The surgical phases are divided into three classes, namely, "Incision", "Capsulorhexis" and "Other". On identifying a phase as "Incision", the incision knife and cornea are segmented whereas if the phase is classified as "Capsulorhexis", only the cornea is segmented. This data will be used to find the horizontal depth of the incision in the incision phase. Additionally, the main contribution of this research is a guidance map generated for the incision and capsulorhexis phases to help surgeons. It also aims to optimize the surgical procedure and reduce the learning curve for a young practitioner. The Incision Phase consists of side port incisions and a main incision for which the incision points are displayed. Along with the incision points, the depth of the main incision along a horizontal plane is calculated. Capsulorhexis is a manoeuvre in cataract surgery where the surgeon tears the anterior capsule of the lens to gain access to the lens cortex and nucleus [2]. In this phase, the rhexis circle(the circle where the capsule is torn) is plotted.

II. Related Work

Artificial Intelligence is fast growing and has been adapted to many bio-medical applications recently. The main focus of this study is to integrate data processing, phase recognition and segmentation to develop a guidance map. In [3], the research focuses on introducing a dataset that has 1000 videos of cataract surgery to be used for phase recognition and segmentation. In the paper, they use their dataset to benchmark a few models for both tasks. This robust dataset is used for this research and its benchmarks are utilised to determine the direction of implementation.

Surgical Phase Recognition refers to the use of deep learning models to perform the task of automatic identification and classification of different phases that are performed during a surgical procedure. The application of such a task in real-time enables the doctors to obtain insights, facilitates workflow optimization, provides surgical assistance and helps in training and evaluation. Phase Recognition has long been a central point of interest for researchers. Its possible use for the

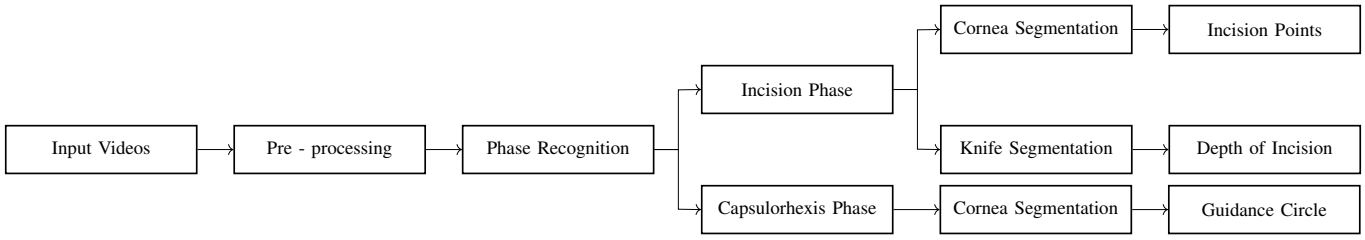


Fig. 1: Flow chart representing methodology of complete application

automatic identification of surgical phases aims to provide actionable insights and meaningful feedback for a complex automated surgical feedback system. Numerous studies, like [3] and [4], focused on using Convolutional Neural Network (CNN) models as backbones paired with Recurrent Neural Network (RNN) frameworks. Multiple other studies such as [5] and [6] also utilised dense and heavy models such as InceptionV3 for the task of surgical phase recognition. The study [7], introduced a new model 'CatStep' for surgical phase recognition which was an ensemble model built of three distinct models, i.e, DenseNet169 for spatial processing of the frames, DenseNet with RNN framework to introduce temporal dependencies in frames and an Inflated-3D CNN (I3D) model to extract spatial-temporal features in frames. Such frameworks, as mentioned above, provide accuracy but usually involve dependency and complexity issues that make these models computationally expensive, and thus not appropriate to be used for real-time predictions.

To perform phase recognition in real-time, authors choose to resort to a highly efficient model, with a less dense and complex architecture while ensuring that it has good feature extraction and a very short inference time. As discussed in the study [8], ResNet-18 is one such architecture. As evidenced in the studies [3], [4], and [9], ResNet-18 as backbone with RNN as top model, outperformed any combination of CNNs.

Semantic segmentation is defined as the classification of pixels into different classes using Deep Learning (DL) methods [10]. This task is crucial and has been researched extensively to adapt to the high accuracy required in medical applications. Due to their ability to effectively learn appearance features from labeled training datasets, DL techniques became popular in medical image analysis [11]. The study in [3] introduced a robust dataset and implemented many segmentation models. The best performance corresponded to the DeepPyramid network with Visual Geometry Group (VGG) 16 as backbone which consistently yielded optimal results in all classes. In [12], UNet could segment large anatomical structures but faced difficulty to segment instruments. This was overcome using UperNet [13], a combination of UNet with a pyramid pooling module which could segment finer features like instruments along with anatomical features. Further, [14] shows a significant improvement in using VGG19 as the backbone in place of VGG16.

There are several other ophthalmology applications that use various segmentation methods. In [15], the segmentation of fine retinal vessels from fundus images was performed. Fine retinal vessels are comparable to instruments in a surgical scene. A guidance filter module was introduced to the UNet model to counter the loss of information occurring due to the downsampling in UNet. The addition of GFM increased the parameters of UNet by only 0.418M and resulted in better performance compared to only UNet. UNet is a model that has been tailored for biomedical image segmentation tasks due to the extreme attention it has received from researchers. This research caused the addition of extensions that enhance the UNet backbone as stated in [16].

A feasible loss function for cataract image segmentation with UNet backbone was concluded to be Cross entropy with Dice from research in [17]. For metrics used to compare various models, IoU and Dice scores have been accurately classified to be F-measure based metrics and are most widely used to measure performance in the medical image segmentation field [18].

You Only Look Once (YOLO) is a well-known model commonly used for segmentation in real-time applications. The YOLOv8 model performs faster than previous versions as shown in [19] and due to fewer parameters it is ideal for real-time applications. It also mentions the lower speed of higher versions despite higher segmentation accuracy, hence YOLOv8 offers optimal speed-accuracy trade-off. This is supported in [20] and the results of YOLOv8 is a benchmark for the inference times of real-time segmentation models.

III. Methodology

The Fig. 1 represents the methodology followed to implement the proposed application. The methodology mainly consists of data pre-processing functions followed by phase recognition and semantic segmentation. This extracted information from the Capsulorhexis Phase is used for cornea segmentation to generate a rhexis circle. Data extracted from the Incision Phase is then processed and used in cornea segmentation to map the incision points, parallelly the data is utilized for knife segmentation to find the horizontal depth of incision.

A. Dataset and Pre-processing

The dataset used is obtained from [3], titled 'Cataract - 1K'. As the name suggests it consists of 1000 cataract surgery videos. The surgeries were performed in the eye clinic of Klinikum Klagenfurt from 2021 to 2023. The surgeons' experience ranged from 1000 to 40,000 surgeries.

For the surgical phase recognition procedure, frames were collected at 30 FPS from 56 video recordings. The video frames are provided with frame-level annotation corresponding to the phase it belongs to. These videos have a spatial resolution of 1024 x 768. The average duration of these videos is 6.54 minutes with a standard deviation of 2.04 minutes. The age of the patients in the videos ranges from 51 to 93 years with an average of 75 years and a standard deviation of 8.69 years. The average experience of the surgeons in these videos is 8929 surgeries with a standard deviation of 6350 surgeries.

The frames extracted consisted of three classes, namely, "Capsulorhexis Phase", "Incision Phase" and "Other". The "Other" class consisted of frames belonging to the remaining surgical phases. Frames obtained from the video are further processed by applying multiple random augmentations aimed at increasing the diversity of the dataset and further enhancing the generalization of the model. These augmentations were decided based on the results obtained by the study [21], which highlighted augmentations such as rotation, translation, and brightness as some of the most effective techniques of augmentations to increase diversity. Input images to the model are resized to 224x224 dimensions to satisfy the input specifications of the model.

The following augmentations were applied to randomly selected images: rotation [-15°, 15°], translation [X, Y]: [-0.05, 0.05], brightness [-20, 20], hue and saturation [-10, 10]. Table I displays the split of the 56 surgical videos across the training, validation and testing sets. Out of the total pool of videos, 42 videos were randomly selected for model training purposes. Of the remaining videos, 7 are randomly selected to be the part of the validation set and the remaining are used for testing.

TABLE I: Table depicting dataset split for Phase Recognition

Dataset Split	Number of Videos
Train	42
Validation	7
Test	7
Total	**56**

For the segmentation procedure, a total of 30 videos were utilized. The annotations provided by [3] are used to make masks for extracted frames. The frames and masks are then segregated into the cornea and the main incision knife for training various models. The cornea masks were obtained for 1713 frames and 15 masks for incision knife. The cornea segmentation dataset has adequate data while the incision knife has a dearth. To compensate for the lack of data, augmentations such as rotation, shear transformation and translation were used to increase it to 240 frames. Further, the details of the augmentations applied are shown in Table II. Augmentations applied were as follows (a) Translation [X, Y]: [0, 0.1] (b) Shear [X, Y]: [-0.3, 0.3] (c) Rotation: [-30°, -25°, -20°, -15°, -10°, -5°, 5°, 10°, 15°, 20°, 25°, 30°]. The training, validation and test sets contain 192, 32 and 16 pairs respectively. The cornea segmentation dataset consisted of 1713 frames and their masks, which were split into 1559 frames, 80 frames, and 74 frames for training, validation, and testing, respectively.

TABLE II: Augmentations applied on the segmentation dataset and their values

Augmentation	Value
Translation-X, Translation-Y	[0,0.1]
Shear-X, Shear-Y	[-0.3,0.3]
Rotation	-30°, -25°, -20°, -15°, -10°, -5°, 5°, 10°, 15°, 20°, 25°, 30°

B. Surgical Phase Recognition

This section focuses on accurately classifying the video frames into three classes, namely Capsulorhexis, Incision, and Other. The decision to undertake only 3-class classification was inspired by the study [5] which stated that 3-class classification is simpler and less computationally heavy as compared to earlier studies that focused on 8, 10, and 14 class classification. For achieving the classification task, ResNet-18 architecture is employed. The decision to implement ResNet-18 architecture was taken after studying the core functionality of ResNet topologies, particularly the ResNet-18 model. The model is lightweight yet effective, with great feature extraction capabilities. The operation of skip connections tackles the problem of vanishing gradients which has been a recurring problem in several of the deep and intricate architectures proposed in previous studies.

Based on careful analysis and experimentation, ResNet-18 is the most promising model for real-time surgical phase recognition. It occupies a space where simplicity, robustness, and efficacy intersect: an ideal solution to address the shortcomings of past methodologies while at the same time providing real-time actionable insight towards surgical skill assessment and feedback.

ResNet-18 is a CNN based on residual networks and skip connections to mend the vanishing gradient problem. The model has 18 layers, which comprises 17 convolutional layers and 1 fully connected layer. Each convolution layer makes use of a 3x3 kernel, with the output feature map size rising as we go deeper into the network. The architecture contains an input block followed by a convolution layer. The model consists of 4 distinct blocks and each block contains 4 convolution layers within.

TABLE III: ResNet-18 Architecture

Stage	Layer
Conv1	Convolution Layer
Block 1	[3x3 , 64] x 4 Conv Layers
Block 2	[3x3 , 128] x 4 Conv Layers
Block 3	[3x3 , 256] x 4 Conv Layers
Block 4	[3x3 , 512] x 4 Conv Layers
FC	Fully Connected Layer

Table III displays the kernel and varying filter dimensions of the convolution layers in the multiple blocks and stages

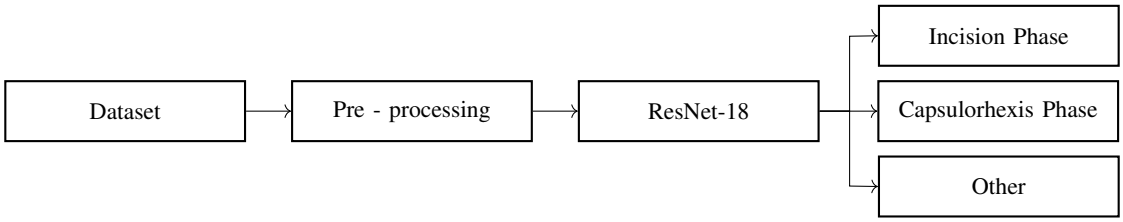


Fig. 2: Flowchart representing phase recognition methodology

of the model. The skip connections are a fundamental feature of ResNet-18 Architecture, that enables the model to combat the problem of vanishing gradients as mentioned earlier. Skip connections allow the outputs from one layer to act as an input to a deeper layer while skipping certain middle units. Hence, the model can learn the identity mapping functions.

Fig. 2 depicts the methodology and the flow of steps followed to achieve the task of phase recognition. In the initial steps, a series of pre-processing functions such as augmentation and resizing are performed on the dataset to make it more diverse and well-suited for the training of the model. Post that, the dataset is passed through the ResNet-18 model to perform the task of frame-level classification in which each video frame received by the model is tagged with the most appropriate surgical phase, i.e, Capsulorhexis Phase, Incision Phase or Other Phase. For the purpose of training, GPU P100 was utilized. Adam Optimizer, with a learning rate of 0.01 was the choice of optimizer. To schedule the learning rate, CyclicLR was employed with a base learning rate of .0001 and a max learning rate of 0.01 in the 'triangular2' mode.

C. Segmentation

The scope of this paper defines two main features to be extracted from the frames of the surgical videos: the main incision knife and the cornea. The main aim is to select a model with a good trade-off between accuracy and speed for real-time processing. The models chosen for segmentation are DeepPyramid+, Unet with GFM and YOLOv8. The DeepPyramid+ with VGG16 is implemented as a benchmark for accuracy and YOLOv8 for inference time, to compare and determine the most suitable real-time segmentation model. UNet with GFM is a trade-off between the two models based on this research. It is also noted that DeepPyramid+ is a heavy model when compared to the other two models requiring more training time.

The segmentation abilities of the chosen models for cornea extraction were tested first, followed by the main incision knife extraction. Since DeepPyramid+ VGG16 and VGG19 performed similarly, only three models were trained for incision knife segmentation.

The steps involved in segmentation are explained in the Fig. 3. The methodology starts with application of pre-processing functions on the dataset. Post that, the pre-processed data is parallelly passed to the DeepPyramid+ VGG16, DeepPyramid+ VGG19, UNet + GFM and YOLOv8 models to perform the task of segmentation. Once the task is completed, the outputs generated by the models are compared and the best one is chosen to be included for our application.

1) Incision Knife Segmentation: DeepPyramid+ is a network that adopts UNet architecture at its core, with the encoder part set to VGG16. The DeepPyramid network is enhanced by adding Pyramid View Fusion and Deformable Pyramid Reception modules [22]. The Pyramid View Fusion block enhances the relative information at each pixel position by replicating the deduction process of the human visual system. The Deformable Pyramid Reception block enhances the accuracy and robustness of the model by handling deformable shapes and heterogenous classes through adaptive feature extraction. DeepPyramid+ is an extension of DeepPyramid with minor enhancements to the DPR module. The implementation of the network was directly taken from [23].

The U-Net architecture is a fully convolutional neural network designed with an encoder-decoder structure and enhanced by skip connections. The encoder path extracts semantic and contextual features by progressively downsampling the input through convolutional layers and pooling operations. The decoder path reconstructs the feature maps to the original resolution using upsampling and transverse convolutional layers for precise pixel-wise segmentation. The defining aspect of U-Net is its skip connections, which transfer feature maps from each stage of the encoder path to the corresponding stage in the decoder path. These skip connections are important to perform biomedical image segmentation [24]. This model was modified by adding a GFM at the first encoder block to improve the segmentation of finer details and provide sharper segmentation outlines. Image filters like the guided image filter [25] are crucial in medical image processing [26]. In GFM by [15], the guided image filter is applied to the feature map to integrate additional fine structural information from the guidance image. The guidance image is chosen to be the grayscale version of the original image. Specifically, the output of the GFM, O, is concatenated with the original feature map, F, to create a more robust representation, as F may retain fine structures that the GFM might have filtered. The algorithm for the GFM is implemented as in [15]. The UNet architecture used is similar to that in [27]. The proposed modified version consists of a single GFM rather than multiple as presented by [15]. The modified version is shown in Fig. 4, where F represents the Feature Map and G represents the Guidance Image. The output O is obtained from the 2D convolution layer.

The Fig. 5 displays the architecture of UNet with GFM

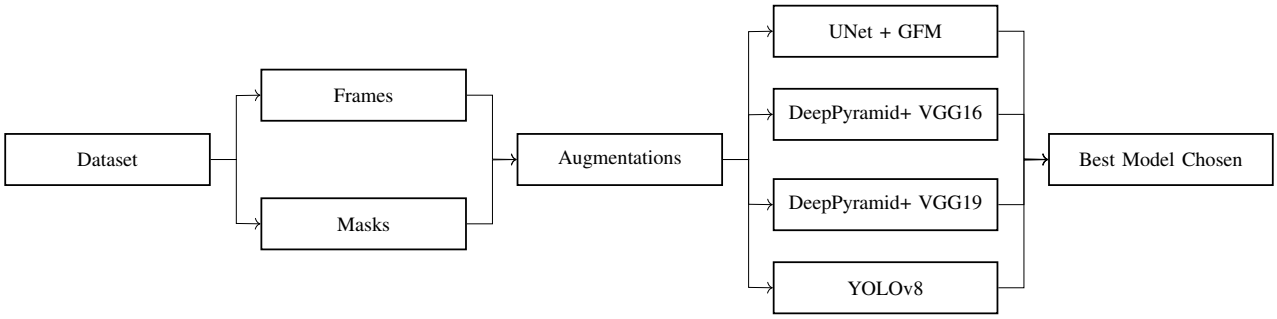


Fig. 3: Flowchart representing segmentation training

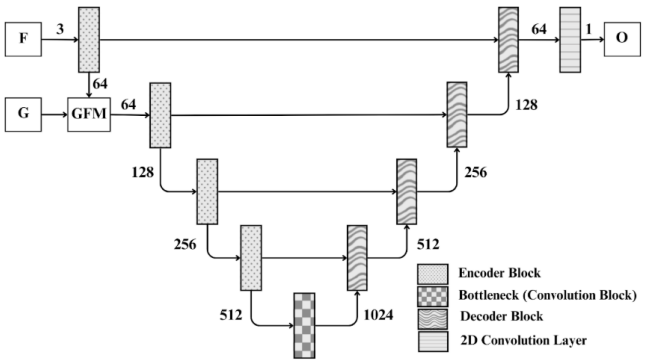


Fig. 4: Architecture of UNet with GFM.

displaying the encoder blocks, decoder blocks along with the 2D convolution layers and the bottleneck. The inclusion of only one GFM reduces the model size and reduces inference time which is crucial in real-time applications.

A YOLOv8 model which has been pre-trained on COCO val2017 [19] was further improved to segment the incision knife by training it on the dataset. It is a real-time object detection and segmentation model that processes input images through a backbone for feature extraction, aggregates multi-scale features using a neck and predicts bounding boxes, class labels, and confidence scores via an anchor-free head. It uses Non-Maximum Suppression (NMS) to refine detections and gives accurate results, optimized for speed and efficiency in various applications. The implementation of this model is from [19].

2) Cornea Segmentation: The DeepPyramid+, UNet with GFM and YOLOv8 are implemented similarly as mentioned in the previous sub-section. The DeepPyramid+ model was trained with VGG16 as the backbone, and the implementation was provided by [23]. The backbone was then changed to VGG19, a newer model, to improve segmentation keeping the rest of the model as in [23] to observe the changes.

All models were given an input image of size 512x512. They were trained for 15 epochs using L4 GPU available on Google Colaboratory. The metrics used to evaluate the models were Mean IoU along with Mean Dice score. Inference time was also noted.

The other training parameters for DeepPyramid+ are a learning rate of 0.005 with the loss function and optimizer being DiceBCE and Stochastic Gradient Descent respectively. Similarly, for UNet with GFM, the learning rate was 0.0001 with the loss function and optimizer being DiceBCE and ADAM respectively. YOLOv8 had the training parameters to be a patience of 0, a confidence threshold of 0.2, and a training batch size of 4.

D. Guidance Map

The work done in this study concludes with the plotting of the guidance map. In order to plot the guidance map, the inputs required are the original frame, the phase it belongs to and its respective segmentation masks. Another important factor required to plot the guidance map is the conversion factor for the number of pixels to a standardized unit. For this reason, we have assumed the diameter of the limbus to be a constant of 11mm, after consulting an ophthalmologist, where the limbus can be defined as the transitional zone or border between the cornea and the sclera. The guidance map is plotted differently for the incision and capsulorhexis phases, which will be explained further in this section.

The algorithm for plotting the guidance map is given in Algorithm 1. C_c refers to the contour of the cornea, $center$ and r refer to the center and the radius of C_c, mm_pp refers to the constant used to convert the number of pixels to millimeters, C_k refers to the contour of the incision knife, e refers to the center of the ellipse that is fit onto the incision knife mask, e_c refers to the contour of the ellipse, θ refers to the incline of the ellipse, P is an array of projections along with its corresponding points, $p[0]$ refers to the projection value in p, (x_0, y_0) refers to $center$, (x_1, y_1) refers to p_{min}, (x_2, y_2) refers to p_{max}, d refers to the distance between the widest segment of the incision knife to $center$, d_{mm} refers to d in millimeter, Δ refers to the horizontal depth of incision, a_L and a_R refer to the angles of the left and right side ports. A line over the widest part of the incision knife can be defined as the line between the points p_{min} and p_{max}.

1) Incision: For the incision phase, the main incision point and the side ports must be plotted. The surgeon is given an option to choose the angle at which the main incision must be done, while the usual range lies between 30° to 45° as shown in Fig. 5. The side ports are 90° and 60° away from the main incision point. The position of the side ports is dependent on

Algorithm 1 Guidance Map Algorithm

Require: `cornea_mask`,`original_img`,`phase`,`knife_mask`, `main_incision_angle`,`hand`

$C_c \leftarrow \max(\texttt{FIND\_CONTOURS}(\texttt{cornea\_mask}))$
$(center, r) \leftarrow \texttt{MIN\_ENCLOSING\_CIRCLE}(C_c)$
$mm\_pp \leftarrow 5.5/r$
`PLOT(CIRCLE(original_img, center, r))`
if `phase` = "Incision" **then**
  $C_k \leftarrow \max(\texttt{FIND\_CONTOURS}(\texttt{knife\_mask}))$
  $(e, e_c, \theta) \leftarrow \texttt{FIT\_ELLIPSE}(C_k)$
  $P \leftarrow []$
  for all $pt \in C_k$ **do**
    $x' \leftarrow (pt_x - e_{c,x})\cos\theta + (pt_y - e_{c,y})\sin\theta$
    Append (x', pt) to P
  $p_{min} \leftarrow \arg\min_{p \in P}(p[0])[1]$

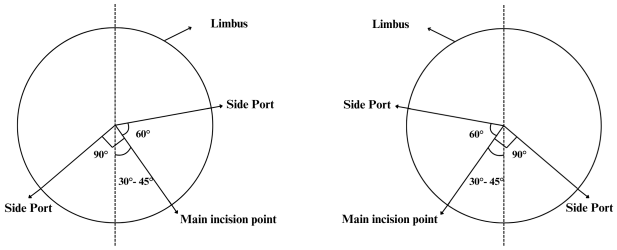
  $d \leftarrow \frac{|(y_2-y_1)x_0-(x_2-x_1)y_0+x_2y_1-y_2x_1|}{\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}}$

  $\Delta \leftarrow d_{mm} - 5.5$
  if $\Delta > 0$ **then**
    `PRINT`(Δ + "mm outside")
    if $1 \leq |\Delta| \leq 1.5$ **then**
      `PRINT`("Correct range")
    else
      `PRINT`("Wrong range")
  else
    `PRINT`($|\Delta|$ + "mm inside")
  if `hand` = "right" **then**
    $a_L \leftarrow main\_incision\_angle - 90 + 360$
    $a_R \leftarrow main\_incision\_angle + 60$
  else
    $a_R \leftarrow 90 - main\_incision\_angle$
    $main\_incision\_angle \leftarrow 360 - main\_incision\_angle$
    $a_L \leftarrow main\_incision\_angle - 60$
  `DRAW_INCLINED_ARROW(original_img, center, r, main_incision_angle)`
  `DRAW_INCLINED_ARROW(original_img, center, r, a_L)`
  `DRAW_INCLINED_ARROW(original_img, center, r, a_R)`
else
  `PLOT(CIRCLE(original_img, center, r/2))`

the dexterity of the doctor which will also be taken as input. The depth of the incision knife along a horizontal plane is also calculated and displayed for each frame as shown in Fig. 6. This distance is calculated between the widest segment of the instrument and the limbus, using Eq 1. For a good incision, this distance must be between 1 to 1.5 mm, as advised by the doctor. In order to find the widest segment of the main incision knife, an ellipse is fitted onto its contour and each point on the contour is projected onto the minor axis using Eq 2. Taking the minimum and maximum of these projections gives us the points p_{min} and p_{max}.

$$d = \frac{|(y_2 - y_1) \times x_0 - (x_2 - x_1) \times y_0 + x_2 \times y_1 - y_2 \times x_1|}{\sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}} \tag{1}$$

$$\text{projection} = (x - x_c) \times \cos(\theta) + (y - y_c) \times \sin(\theta) \tag{2}$$



(a) (b)

Fig. 5: Incision guidance map for dexterity: (a) Right-handed and (b) Left-handed.

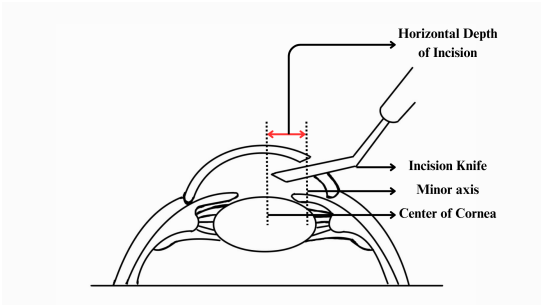


Fig. 6: Horizontal depth of incision.

2) Capsulorhexis: For the capsulorhexis phase, the main aim is to plot the rhexis circle, which the doctor can follow in order to make the tear in the capsule. The rhexis circle should ideally be 5.5 mm in diameter.

IV. Results and Discussion

This section presents a detailed analysis of the performance of ResNet-18 in the phase recognition task, along with the insights captured by the various models for the semantic segmentation task. It also shines light on the guidance map that helps in enhancing the understanding of surgical tasks by plotting crucial incision points, side ports and the horizontal depth of the instrument.Further, it discusses the plotting of the rhexis circle during the capsulorhexis phase providing the necessary guidance to surgeons.

A. Phase Recognition

The performance of ResNet-18 on the surgical phase recognition task was found to be exceptionally good, achieving high accuracy and low inference time. The model is evaluated based on a wide range of metrics, such as accuracy and F1 scores. Other classification metrics include, "precision", which reflects the number of true positive predictions of total positive predictions made, "recall", showing the proportion of actual positives from the predicted positives, "specificity", which gives the true negatives out of the total real negatives available for the computation, and "inference time", which reflects the time taken by the model to make a prediction.

Table IV shows the evaluation metrics employed to evaluate model performance. These metrics are taken from the test data

(a) (b) (c) (d) (e)

Fig. 7: Outputs of Incision Knife Segmentation (a) Image (b) Ground truth (c) DeepPyramid+ VGG16 predicted mask (d) UNet with GFM (47 epochs) predicted mask (e) YOLOv8 predicted mask.

(a) (b) (c) (d) (e) (f)

Fig. 8: Outputs of Cornea Segmentation (a) Image (b) Ground truth (c) DeepPramid+ VGG16 predicted mask (d) DeepPyramid+ VGG19 predicted mask (e) UNet with GFM (35 epochs) predicted mask (f) YOLOv8 predicted mask.

that the model has never seen before. Among these metrics, the model achieved a test accuracy of 0.9442 and a test F1 score of 0.9452. The model attained high precision for the Capsulorhexis Phase with the score of 0.98754, a high recall value for Other Phase and great specificity score of 0.96192 for the Incision Phase. The model also resulted in a very fast inference time of 7.3ms, 2.7ms, and 3.2ms for Capsulorhexis Phase, Incision Phase, and Other Phase respectively.

TABLE IV: Evaluation Metrics for ResNet-18 Architecture

Metric	Capsulorhexis	Incision	Other
Precision	0.98754	0.95394	0.87736
Recall	0.9179	0.97252	0.97428
Specificity	0.95096	0.96192	0.92086
Inference Time (s)	0.0073	0.0027	0.0032

B. Segmentation

The overall performance of UNet with GFM, considering the accuracy-speed trade-off, was adequate for the proposed application. The mean IoU and mean Dice scores were found to be comparable to the best-performing model while its inference time was found to be better than DeepPyramid+ and YOLOv8. The formulae for the metrics are given in Eq. 3 and 4.

$$meanIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{Intersection_i + \epsilon}{Union_i + \epsilon}. \tag{3}$$

$$meanDice = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \cdot Intersection_i + \epsilon}{Total_i + \epsilon}. \tag{4}$$

In Eq. 3 and 4, the intersection is calculated as $Intersection = \sum(inputs \cdot targets)$, and the union is calculated as $Union = \sum(inputs + targets) - Intersection$. The total can be defined as $Total = \sum(inputs + targets)$. ϵ is added in the equations for numerical stability to avoid division by zero and N refers to the batch size. $Intersection_i$ and $Union_i$ are calculated for the i^{th} sample in the batch in terms of area. *inputs* refer to the predicted masks and *targets* refer to the ground truth masks. The inferences were made on Google Colaboratory using the L4 GPU. The results obtained will be discussed next.

1) Incision Knife Segmentation: The below set of images displays the output produced by the various different segmentation models which are mentioned earlier in the methodology. The test images and ground truth masks are shown in Fig. 7(a) and Fig. 7(b) respectively. The corresponding predicted masks segmented using various models are shown in Fig. 7(c) displaying the output of the DeepPyramid+ VGG16, Fig. 7(d) displaying the output of the UNet with GFM (47 epochs) and Fig. 7(e) displaying the output of the YOLOv8.

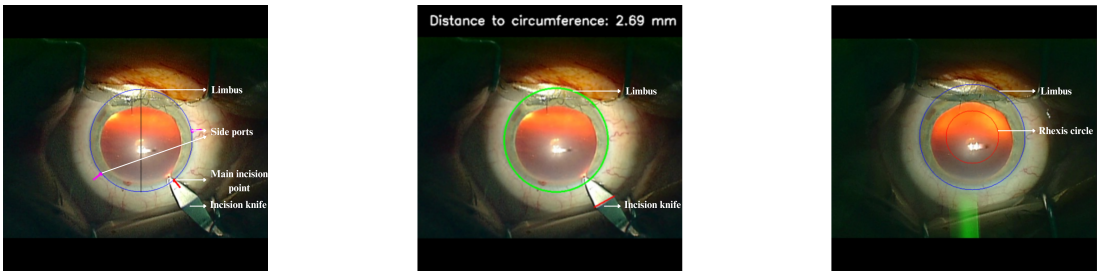
The metrics obtained by implementing the models specified in the methodology are given in Table V. It shows that the inference time for DeepPyramid+ VGG16, YOLOv8 and UNet with GFM are 9.4ms, 55.8ms and 6.8ms, respectively when trained for 15 epochs. Among these models, the mean IoU and mean Dice scores were observed to be comparable. The DeepPyramid+ VGG16 model produced masks with mean IoU and mean Dice scores of 0.9122 and 0.9539. Similarly, YOLOv8 and UNet with GFM produced masks with mean IoU scores of 0.9307 and 0.9238 and mean Dice scores of 0.9641 and 0.9601. On further training the UNet with GFM model, a mean IoU score of 0.9315 and a mean Dice score of

TABLE V: Metrics of Incision Knife Segmentation

Metric	DeepPyramid+ VGG16	YOLOv8	UNet with GFM	UNet with GFM (Best of 50 epochs)
Epochs	15	15	15	47
Mean IoU	0.9122	0.9307	0.9238	**0.9315**
Mean Dice	0.9539	0.9641	0.9601	**0.9642**
Inference time in s	0.0094	0.0558	0.0068	**0.0062**

TABLE VI: Metrics of Cornea Segmentation

Metric	DeepPyramid+ VGG16	DeepPyramid+ VGG19	YOLOv8	UNet with GFM	UNet with GFM (Best of 50 epochs)
Epochs	15	15	15	15	35
Mean IoU	**0.9455**	**0.9455**	0.9411	0.9436	0.9450
Mean Dice	**0.9719**	0.9718	0.9696	0.9709	0.9716
Inference time in s	0.0095	0.0097	0.0425	0.0066	**0.0065**



(a) (b) (c)

Fig. 9: Result of guidance map (a) Incision phase (b) Incision depth tracking (c) Capsulorhexis phase

0.9642 were obtained along with an inference time of 6.2ms.

It can be inferred from Table V that the UNet with GFM model achieved an inference time of 6.2ms which is lesser than YOLOv8 which had an inference time of 55.8ms. The inference time has been considered as our primary metric to conclude on our best model as the mean IoU and mean Dice scores are comparable. On training the UNet model for 50 epochs, the best model was obtained at the 47th epoch with the maximum mean IoU and mean Dice scores. The UNet with GFM model gives an adequate mean Dice and mean IoU score for the proposed application and thus training was terminated at 47 epochs.

2) Cornea Segmentation: The test images and ground truth mask are shown in Fig. 8(a) and Fig. 8(b) respectively. The above set of images displays the output produced by the various different segmentation models which are performing the task of cornea segmentation. Fig. 8(c) displays the output of the DeepPyramid+ VGG16, Fig. 8(d) displays the output of the UNet with GFM (35 epochs) and Fig. 8(e) displays the output of YOLOv8.

The metrics obtained on implementing the models specified in the methodology are given in Table VI. It shows that the DeepPyramid+ VGG16 model obtained results with mean IoU and mean Dice scores of 0.9455 and 0.9719 respectively with inference time of 9.5ms. This is comparable to the values obtained for the DeepPyramid+ VGG19 model that obtained values of 0.9455, 0.9718 and 9.7ms for mean IoU, mean Dice and inference time respectively. YOLOv8 obtained masks with mean IoU, mean Dice and inference time of 0.9411, 0.9696 and 42.5ms respectively. UNet with GFM gave scores of 0.9436 and 9709 for mean IoU and mean Dice when trained for 15 epochs, equal to the other models. When trained further, it obtained values of 0.9450 and 0.9716 for mean IoU and mean Dice. However, the inference time was reduced from 6.6ms to 6.5ms when trained further above 15 epochs.

It can be inferred from Table VI that the inference time for the UNet with the GFM model is the least at 6.5ms. On training the UNet Model for 50 epochs, the mean IoU and mean Dice scores are comparable to DeepPyramid+ at the 35th epoch with the same level of accuracy scores being achieved. The Mean IoU and Dice score obtained are adequate for the proposed application and hence 35 epochs were sufficient.

C. Guidance Map

In Fig. 9(a) and Fig. 9(b), we can see the output frame after the guidance map was plotted on the original image for the incision phase. For the incision phase, the main incision point and the side ports are plotted and the horizontal depth from the widest segment of the instrument is displayed on each frame. The output obtained is a video where frames are combined at 30FPS. The terminal also displays whether the incision knife is within range or out of range. For the capsulorhexis phase,

the rhexis circle with a diameter of 5.5 mm is plotted as shown in Fig. 9(c).

V. Conclusion and Future Scope

This paper integrates phase recognition, semantic segmentation and guidance map plotting. With respect to phase recognition, ResNet-18 was able to achieve high performance metrics while having a very low inference time making it suitable for real-time application. For the task of cornea segmentation, the DeepPyramid+ models performed the best in terms of Dice and IoU but UNet with GFM gave better inference time with comparable accuracy scores. UNet + GFM gave the best accuracy scores along with lesser inference time for incision segmentation. Hence, UNet with GFM was used for both the segmentation tasks.

For this study, we focused on plotting the guidance map for only the "Incision" and "Capsulorhexis" phases. As a future scope, other phases of cataract surgery and addressing irregularities occurring during the surgery can be considered. A larger and more diverse dataset can be used to train the model making it better generalized and more robust. As an extension, the project can be converted to a full-scale, end-to-end application that is compatible with currently existing microscopes used to perform cataract surgeries. As the world moves towards automation, this research can be considered the first step towards robotic cataract surgeries.

Acknowledgements

We express our sincere thanks to PES University for providing the essential resources and steadfast support that made this study possible.

References

[1] Xinyi Chen, Jingjie Xu, Xiangjun Chen, Ke Yao, Cataract: Advances in surgery and whether surgery remains the only treatment in future, Advances in Ophthalmology Practice and Research, Volume 1, Issue 1, 2021, 100008, ISSN 2667-3762. DOI: https://doi.org/10.1016/j.aopr.2021.100008

[2] Nathan Pan-Doh, Shameema Sikder, Fasika A. Woreta, James T. Handa, Using the language of surgery to enhance ophthalmology surgical education, Surgery Open Science, Volume 14, 2023, Pages 52-59, ISSN 2589-8450, DOI: https://doi.org/10.1016/j.sopen.2023.07.002

[3] Ghamsarian, N., El-Shabrawi, Y., Nasirihaghighi, S. et al. Cataract-1K Dataset for Deep-Learning-Assisted Analysis of Cataract Surgery Videos. Sci Data 11, 373 (2024). DOI: https://doi.org/10.1038/s41597-024-03193-4.

[4] Bansod, V. and Ambhaikar, A., 2024. Surgical Phase Recognition Using Videos: Deep Neural Network Approach. In Handbook of Research on Artificial Intelligence and Soft Computing Techniques in Personalized Healthcare Services (pp. 189-208). Apple Academic Press.

[5] Morita Shoji, Tabuchi Hitoshi, Masumoto Hiroki, Yamauchi Tomofusa, Kamiura Naotake. Real-Time Extraction of Important Surgical Phases in Cataract Surgery Videos. Sci Rep 9, 16590 (2019). DOI: https://doi.org/10.1038/s41598-019-53091-8

[6] Tabuchi H, Masumoto H, Tanabe H, Kamiura N. Real-Time Surgical Problem Detection and Instrument Tracking in Cataract Surgery. J Clin Med. 2020 Nov 30;9(12):3896. doi: 10.3390/jcm9123896. PMID: 33266345; PMCID: PMC7759772

[7] Mahmoud, O., Zhang, H., Matton, N., Mian, S.I., Tannen, B. and Nallasamy, N., 2024. CatStep: Automated Cataract Surgical Phase Classification and Boundary Segmentation Leveraging Inflated 3D-Convolutional Neural Network Architectures and BigCat. Ophthalmology Science, 4(1), p.100405

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.

[9] Wang, B., Li, L., Nakashima, Y., Kawasaki, R. and Nagahara, H., 2023. Real-time estimation of the remaining surgery duration for cataract surgery using deep convolutional neural networks and long short-term memory. BMC Medical Informatics and Decision Making, 23(1), p.80.

[10] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, Dec. 2017.

[12] Grammatikopoulou, M., Flouty, E., Kadkhodamohammadi, A., Quellec, G., Chow, A., Nehme, J., Luengo, I. and Stoyanov, D., 2021. CaDIS: Cataract dataset for surgical RGB-image segmentation. Medical Image Analysis, 71, p.102053

[13] Xiao T, Liu Y, Zhou B, Jiang Y, Sun J (2018) Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV), pp 418–434

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556. Accessed: Jan. 25, 2025.

[15] P. Yin, R. Yuan, Y. Cheng and Q. Wu, "Deep Guidance Network for Biomedical Image Segmentation, in IEEE Access, vol. 8, pp. 116106-116116, 2020, doi: 10.1109/ACCESS.2020.3002835.

[16] I. Gandomi, M. Vaziri, M. J. Ahmadi, M. Reyhaneh Hadipour, P. Abdi and H. D. Taghirad, "A Deep Dive Into Capsulorhexis Segmentation: From Dataset Creation to SAM Fine-tuning," 2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, Islamic Republic of, 2023, pp. 675-681, doi: 10.1109/ICRoM60803.2023.10412370.

[17] I. Gandomi et al., "A Deep Dive Into Capsulorhexis Segmentation: From Dataset Creation to SAM Fine-tuning," in 2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM), Tehran, Iran, 2023, pp. 675-681. DOI: https://doi.org/10.1109/ICRoM60803.2023.10412370

[18] D. Muller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," BMC Research Notes, vol. 15, no. 1, pp. 1–8, 2022.

[19] https://docs.ultralytics.com/models/yolov8/#overview as on December 29th 2024, 01:28:56.

[20] N. Nyi Myo, A. Boonkong, K. Khampitak and D. Hormdee, "Real-Time Surgical Instrument Segmentation Analysis Using YOLOv8 With ByteTrack for Laparoscopic Surgery," in IEEE Access, vol. 12, pp. 83091-83103, 2024, doi: 10.1109/ACCESS.2024.3412780.

[21] Ramesh, S., Dall'Alba, D., Gonzalez, C. et al. TRandAugment: temporal random augmentation strategy for surgical activity recognition from videos. Int J CARS 18, 1665–1672 (2023). https://doi.org/10.1007/s11548-023-02864-8.

[22] Ghamsarian N, Taschwer M, Sznitman R, Schoeffmann K (2022) Deeppyramid: Enabling pyramid view and deformable pyramid reception for semantic segmentation in cataract surgery videos. In: Medical image computing and computer assisted intervention—MICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, Proceedings, Part V. Springer, pp 276–286

[23] Ghamsarian Negin, Wolf Sebastian, Zinkernagel Martin, Schoeffmann Klaus, Sznitman Raphael. DeepPyramid+: medical image segmentation using Pyramid View Fusion and Deformable Pyramid Reception. Int J CARS (2024). https://doi.org/10.1007/s11548-023-03046-2

[24] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in Deep Learning and Data Labeling for Medical Applications. Cham, Switzerland: Springer, 2016, pp. 179–187.

[25] K. He, J. Sun, and X. Tang, "Guided image filtering," IEEE Trans. Pattern Anal. Mach. Intell., no. 6, pp. 1397–1409, Jun. 2013.

[26] I. Bankman, Handbook of Medical Image Processing and Analysis. Amsterdam, The Netherlands: Elsevier, 2008.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.