

# Epiretinal Membrane and Instrument Segmentation in Intraoperative Retinal Surgery Video

Author: Saurav Verma, M.S. Information Science

Advisor: Dr. Eungjoo Lee

Lab: Vision, Systems & Intelligence (VSI) Lab — University of Arizona

## Abstract

Epiretinal membrane (ERM) peeling in vitreoretinal surgery requires surgeons to identify thin, low-contrast membranes and manage delicate instrument–tissue interactions under challenging optical conditions. While pixel-level surgical video segmentation has been extensively explored in laparoscopic and robotic abdominal surgery, ophthalmic surgical video—particularly intraoperative retinal surgery—is underrepresented in the literature. This capstone project contributes a pilot benchmark for epiretinal membrane and instrument segmentation from intraoperative retinal surgery video, developed within the Vision, Systems & Intelligence (VSI) Lab at the University of Arizona.

We curate a small but carefully constructed dataset of 300 frames sampled at 1 fps from seven ERM peeling surgeries. Each frame is annotated in Supervisely with pixel-wise labels for epiretinal membrane, intraocular retinal forceps, and endoillumination (“light tool”), and tagged with surgical phases (dye, flap initiation, peeling, completion) for future multi-task modeling. The dataset is challenging: ERM and instrument pixels together occupy far less than 1% of the image, membranes may be faint or unstained, and strong specular highlights and tool overlap are common.

To support reproducible experimentation, we design and implement an end-to-end preprocessing and modeling pipeline. Preprocessing includes frame extraction, structural similarity-based deduplication, blur and quality checks, fixed region-of-interest cropping, and joint resizing of images and masks to 512×512 and 768×768. We train modern convolutional segmentation architectures—DeepLabv3+ and UNet++ with ResNet-101 backbones—using a composite Dice + Cross-Entropy loss. Evaluation goes beyond aggregate accuracy and reports present-only Dice, mmseg-style per-class Dice/IoU, Boundary-F1 at a 3-pixel tolerance, and pixel-share consistency between predictions and ground truth.

On this pilot dataset, DeepLabv3+ at 512×512 resolution provides the strongest overall performance, achieving present-only Dice of approximately 0.77 for ERM, 0.83 for forceps, and 0.76 for the light tool, with ERM Boundary-F1  $\approx$  0.78 and foreground mean Dice/IoU  $\approx$  0.74/0.59. ERM pixel share closely matches ground truth (0.159% vs 0.157%), indicating that the model can recover membrane proportion despite extreme foreground sparsity. UNet++ performs slightly lower, and DeepLabv3+ at 768×768 underperforms 512×512, highlighting the interaction between dataset size, resolution, and thin-structure segmentation. Qualitative panels confirm robust performance on clearly visible membranes and instruments.

Overall, this work demonstrates that pixel-level ERM and instrument segmentation from intraoperative retinal surgery video is technically feasible even with limited data, and establishes a benchmark-style framework that can be extended within the VSI Lab to larger datasets, advanced architectures (e.g., transformers), boundary-aware objectives, and joint segmentation–phase recognition models.

## **Research & Code Availability Notice**

*This capstone project is part of an ongoing research program in the VSI Lab. Exact dataset details, source videos, and full implementation code are intentionally omitted from this public report. The emphasis is on high-level methodology, experimental design, and results.*

# 1. Introduction

Epiretinal membrane (ERM) is a fibrocellular proliferation that forms on the inner surface of the retina and can lead to visual distortion, metamorphopsia, and decreased visual acuity. Surgical peeling of ERM is a common vitreoretinal procedure in which the surgeon uses fine intraocular forceps to grasp and lift the membrane away from the retina. During surgery, the membrane can be extremely thin, partially transparent, and difficult to distinguish from the underlying retina, especially in unstained or faintly stained regions. Accurate perception of membrane boundaries and instrument–tissue interactions is therefore critical for safe and effective surgery.

In parallel, computer vision and machine learning have made substantial progress in surgical video understanding over the past decade. In laparoscopic and robotic abdominal surgery, deep learning models now support tasks such as instrument segmentation, tissue classification, surgical tool detection, and workflow or phase recognition. Several benchmark datasets and challenges—including EndoVis, Cholec80, and multi-tool tracking datasets—have driven this progress by providing standardized tasks and metrics.

By contrast, ophthalmic surgical video remains relatively underexplored. Much of the ophthalmic machine learning literature focuses on static imaging modalities such as fundus photography and optical coherence tomography (OCT). These domains have seen a proliferation of deep learning approaches for disease detection, progression modeling, and retinal layer segmentation. However, intraoperative microscope video during vitreoretinal surgery presents a distinct set of challenges: small field of view, microscope optics, specular reflections, rapid instrument motion, and extremely fine anatomical structures such as the ERM.

## 1.1 Problem Statement

This project addresses the task of pixel-level segmentation of epiretinal membranes and surgical instruments from intraoperative retinal surgery video. Concretely, given a surgical frame captured from a microscope during ERM peeling, the goal is to produce per-pixel labels for:

- Epiretinal membrane (ERM)
- Intraocular retinal forceps
- Endoillumination (light tool)

The segmentation must be robust under:

- Thin, low-contrast membranes, often occupying far less than 1% of image pixels.
- Unstained or faintly stained ERM, where membranes are only subtly visible.
- Tool overlap, including forceps occluding the membrane or casting shadows.
- Illumination variability, including specular highlights and glare.

This task is challenging not only due to the underlying physics and optics of the microscope, but also due to data scarcity. Acquiring expert-annotated surgical frames is time-consuming, and clinical privacy constraints limit how broadly data can be shared.

## 1.2 Goals and Contributions

The overarching goals of this capstone project are:

- **Dataset and benchmark creation**

Curate a small but representative dataset of intraoperative ERM surgery frames, with pixel-level annotations for ERM and instruments, and surgery-level phase tags.

- **Pipeline design and implementation**

Build an end-to-end preprocessing and modeling pipeline that can be reused and extended within the VSI Lab, covering frame extraction, quality filtering, annotation preparation, cropping, resizing, dataset splitting, training, and evaluation.

- **Baseline modeling and analysis**

Train modern convolutional segmentation models (DeepLabv3+ and UNet++) under realistic constraints (extreme class sparsity, small dataset) and analyze performance using appropriate metrics, including boundary-aware measures.

- **Foundational framework for future research**

Provide a benchmark-style foundation that can be extended to larger datasets, more advanced architectures (e.g., transformer-based segmentation), boundary-aware objectives, and multi-task models that combine segmentation with surgical phase recognition.

While this report is written in the neutral “we” voice, the project reflects substantial individual contributions in pipeline design, dataset curation, annotation workflow support, experimental implementation, and communication.

## 2. Related Work

### 2.1 Surgical Video Segmentation

Deep learning for surgical video has gained traction especially in laparoscopic and robotic procedures. EndoNet and related architectures have been applied to phase recognition, tool presence detection, and frame classification in cholecystectomy videos. Robotic instrument segmentation challenges have catalyzed progress in pixel-level segmentation, comparing U-Net variants, fully convolutional networks, and more recent attention-based architectures.

These works highlight several general lessons:

- Data scale matters: High-performing models are typically trained on thousands of labeled frames, often augmented with weak labels or semi-supervised techniques.
- Context and temporal information help: Incorporating temporal smoothing or sequence models can improve robustness to motion and transient occlusions.
- Metrics must reflect task structure: For instrument segmentation, foreground Dice and IoU, as well as instance-level metrics, are critical beyond overall accuracy.

However, most of these efforts focus on abdominal organs and instruments in relatively large fields of view, with pixel-wise objects that occupy a more moderate proportion of the image than thin retinal membranes

### 2.2 Ophthalmic Imaging and ERM

In ophthalmology, machine learning has been widely applied to fundus photographs and OCT scans. Deep learning algorithms have shown strong performance in detecting diabetic retinopathy, age-related macular degeneration, and other retinal pathologies. Several works specifically investigate epiretinal membrane detection and characterization on OCT, often leveraging 2D convolutional networks or 3D volumetric architectures.

There is also emerging interest in augmenting intraoperative retinal surgery through feature tracking, augmented reality overlays, and en-face guidance tools. Recent studies have explored real-time segmentation and tracking of surgical instruments in vitreoretinal surgery, as well as intraoperative augmented-reality systems for membrane peeling and retinal detachment management.

Yet, there remains a gap between these technologies and frame-wise pixel segmentation of ERM on real microscope video. Membrane boundaries are often not sharply defined; peeling alters membrane geometry in real time; and the available video data may be limited to a small number of cases.

## 2.3 Positioning of this work

This project sits at the intersection of surgical video segmentation and ophthalmic imaging:

- It applies deep segmentation models to vitreoretinal surgery video rather than to OCT or fundus images.
- It focuses on thin membranes and instruments with extreme foreground sparsity.
- It prioritizes pipeline design and reproducible evaluation in a small-data setting.

Rather than proposing a novel architecture, the main contributions are:

- A pilot dataset and benchmark protocol for ERM + instrument segmentation from intraoperative video.
- A carefully engineered pipeline that addresses deduplication, blur detection, annotation support, and surgery-level dataset splitting.
- A metric suite that includes boundary-aware evaluation and pixel-share analysis, which are particularly important for thin structures.

## 3. Methods

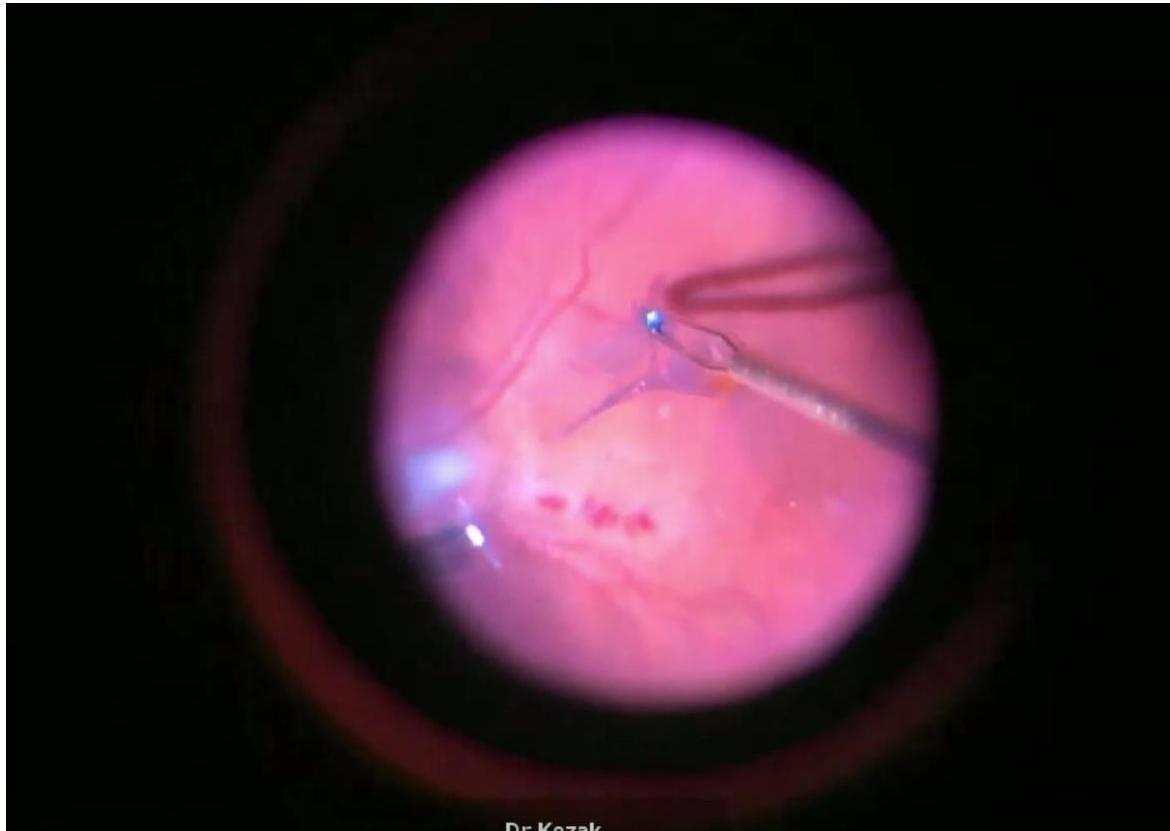
### 3.1 Data Source and Cohort

The dataset comprises intraoperative microscope video from seven ERM peeling surgeries performed at a single center. Videos are captured from the surgical microscope during vitreoretinal surgery. From these videos, we extract frames at 1 frame per second (fps), focusing on segments where ERM peeling and related manipulations occur.

To construct a manageable yet representative pilot dataset, we sample 300 frames across the seven surgeries. Selection emphasizes:

- Coverage of key phases: dye application, flap initiation, active peeling, and completion.
- Inclusion of challenging cases: faint or unstained ERM, overlapping instruments, glare, and strong specular highlights.
- Avoidance of trivial near-duplicates (addressed in preprocessing).

All data remain on secure lab storage; this report describes the dataset conceptually without exposing identifiable or proprietary information.



*Figure 1 Example intraoperative frame showing intraocular forceps initiating an ERM peel.*

### 3.2 Annotation Protocol

Annotations are performed using Supervisely, a web-based annotation platform that supports pixel-wise masks and project-level management. For each frame, we annotate three foreground classes:

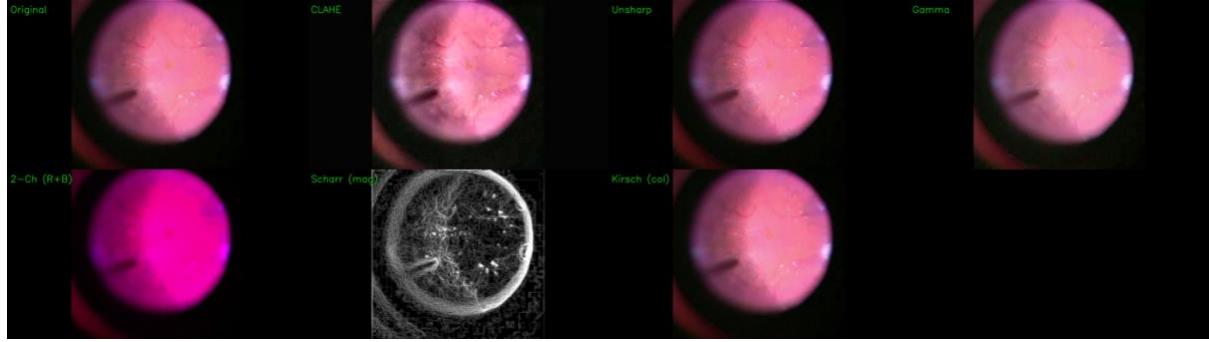
1. Epiretinal membrane (ERM)
2. Retinal forceps
3. Endoillumination tool (“light tool”)

Background pixels are implicitly defined as everything else.

To improve visibility of faint membranes, we generate several enhancement views for annotators:

- Contrast-limited adaptive histogram equalization (CLAHE)
- Unsharp masking and edge-enhanced variants
- Channel-separated or ratio-based views to emphasize membrane boundaries

These enhancement views are not used directly for training but support consistent, higher-quality labeling.



*Figure 2 Image enhancement views are not used directly for training but support consistent labelling*

In addition to pixel-wise masks, each frame is tagged with a surgical phase:

- Dye application
- Flap initiation
- Peeling progression
- Completion

These phase labels are not used in the segmentation models for this capstone but are retained for future multi-task modeling.

Example: Original Frame, Annotation, and Mask

Original Frame	Supervisely Annotation	Sigle channel Mask

*Figure 3 Annotated frame with colored overlays for ERM, forceps, and light tool alongside grayscale mask*

### 3.3 Preprocessing Pipeline

We implement an end-to-end preprocessing pipeline in Python, orchestrated via a configuration file (e.g., config.yaml) and a central script (pipeline.py). The main stages are:

1. Frame Extraction
  - Use ffmpeg to extract frames at 1 fps from raw microscope videos.
  - Store frames in a structured directory layout grouped by surgery.
2. De-duplication via SSIM
  - Compute Structural Similarity Index (SSIM) between neighboring frames.
  - Remove near-duplicate frames above a chosen similarity threshold to avoid over-representing static scenes.
3. Blur and Quality Filtering
  - Compute a variance-of-Laplacian score for each frame.
  - Optionally filter out extremely blurred frames or flag them for review.
4. Region-of-Interest Cropping

- Apply a fixed, surgery-specific ROI mask to crop the circular surgical field.
  - Remove black borders and extraneous regions outside the microscope view.
5. Joint Resizing of Images and Masks
- Resize both images and corresponding masks to  $512 \times 512$  and  $768 \times 768$  pixels.
  - Maintain alignment between images and masks by applying identical transformations.
6. Train/Validation/Test Split
- Perform surgery-level splits to avoid patient leakage:
  - Train: 231 frames
  - Validation: 21 frames
  - Test: 48 frames
  - Ensure that all frames from a given surgery belong to the same split.
7. Training-Ready Dataset
- Normalize images and convert to CHW tensor format for PyTorch DataLoaders.
  - Store metadata (surgery ID, phase tag, presence/absence of each class) for downstream analysis.

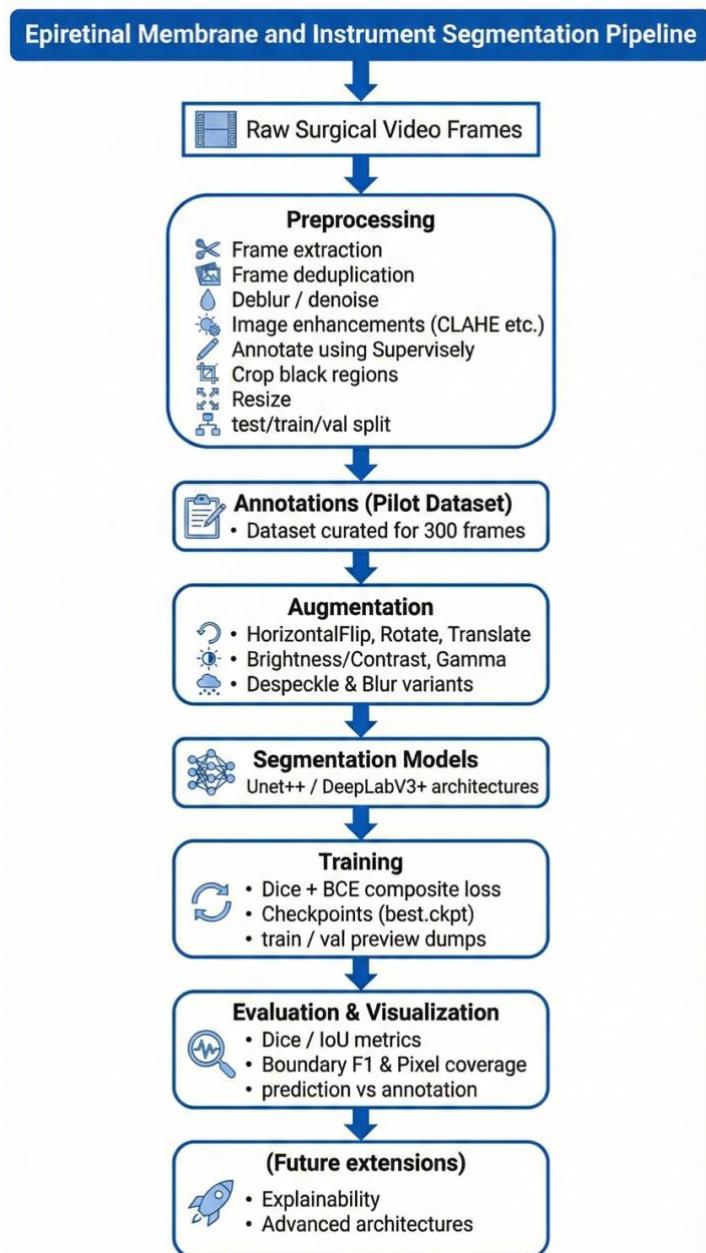


Figure 4 End-to-end ERM surgical segmentation workflow diagram

### 3.4 Data Augmentation

To mitigate overfitting and address the small dataset size, we apply standard spatial and photometric augmentations during training. Augmentations are applied only to the training split and include:

- Horizontal flips
- Small rotations and translations
- Brightness, contrast, and gamma adjustments
- Mild blur and despeckle variants

We visualize augmentation configurations by generating grid panels that show the original image/mask pair and several augmented variants. These panels help verify that transformations preserve membrane and tool geometry in a clinically plausible way.

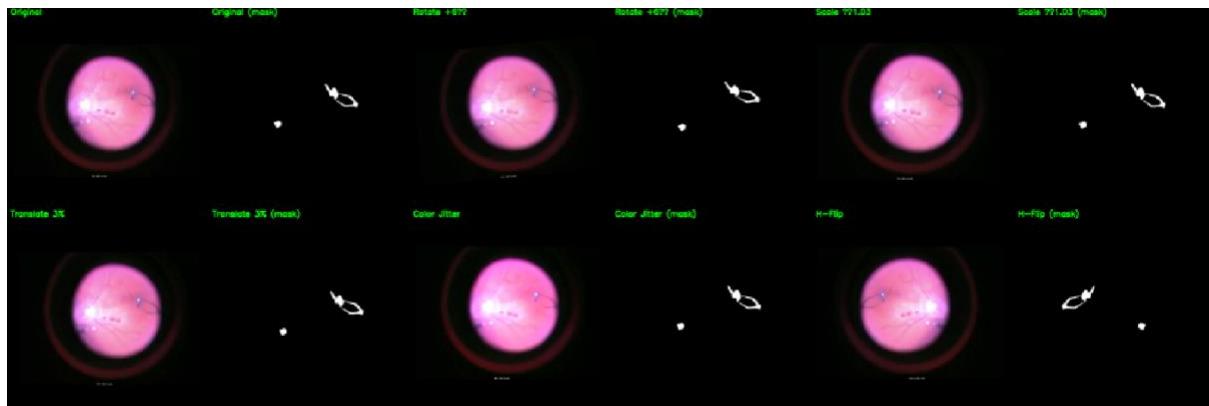


Figure 5 Augmentation grid with original + mask in the top-left and several augmented variants

### 3.5 Model Architecture

We focus on two widely used convolutional segmentation architectures implemented in PyTorch:

#### 1. DeepLabv3+

- Backbone: ResNet-101 encoder pre-initialized on ImageNet.
- Uses atrous spatial pyramid pooling (ASPP) and a decoder head to refine boundaries.
- Well-suited to capturing multi-scale context.

#### 2. UNet++

- Backbone: ResNet-101 encoder with nested skip connections and dense connectivity in the decoder.
- Designed to improve gradient flow and segmentation performance over the original U-Net, especially on medical images.

Both models are trained for multi-class segmentation with the three foreground classes (ERM, forceps, light tool) plus implicit background.

### 3.6 Training setup

Training is implemented in a configuration-driven manner to allow consistent experiments across model types and resolutions.

- Input resolutions
  - 512×512 pixels (primary setting).
  - 768×768 pixels (higher-resolution exploratory setting).
- Loss functions
  - Primary: composite Dice + Cross-Entropy (CE) loss.
  - Experimental: low-weight Tversky loss (e.g.,  $\alpha=0.3$ ,  $\beta=0.7$ ,  $\lambda\approx0.1$ ) added to emphasize recall for thin ERM regions.
  - In practice, the Tversky term did not consistently improve metrics on this small dataset; reported results focus on Dice + CE.
- Optimization and schedule
  - Optimizer: AdamW.
  - Number of epochs: ~80 for pilot runs, with early stopping based on validation loss and Dice.
  - Standard batch size chosen to fit GPU memory (details omitted here as implementation is internal to the lab).
- Data handling
  - DataLoader shuffles training frames and applies augmentations on-the-fly.
  - Validation and test sets are evaluated without augmentation.

### 3.7 Evaluation Metrics

Because ERM and instruments are thin, sparse structures, we emphasize metrics beyond raw pixel accuracy:

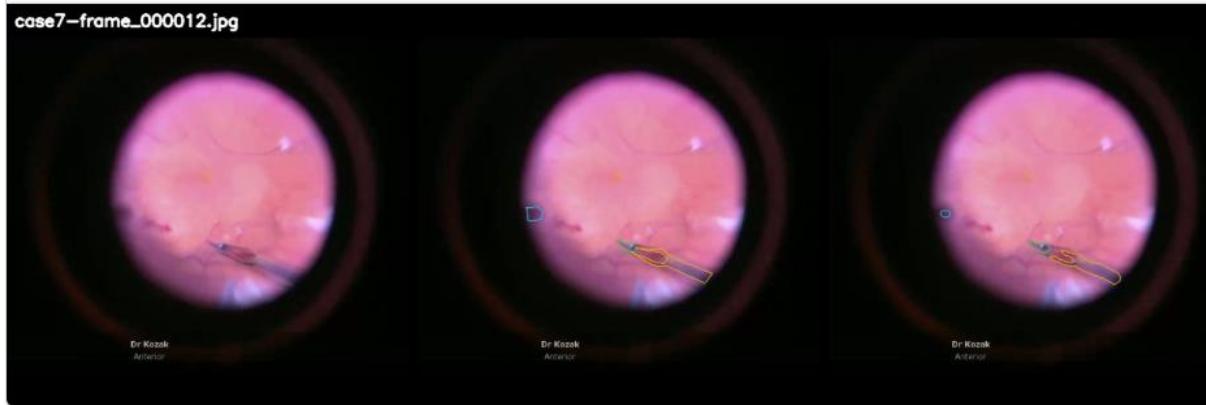
1. Present-Only Dice
  - Soft Dice coefficient computed only on frames where a given class is present in the ground truth.
  - This avoids inflating metrics for classes that are absent in many frames.
2. Overall mean class Dice and IoU
  - Aggregate true positives, false positives, and false negatives across the dataset.
  - Compute per-class Dice and Intersection-over-Union, as well as foreground mean (mDice, mIoU) over the three foreground classes.
3. Boundary-F1 at 3 Pixels
  - Compute F1 score on mask boundaries for ERM with a 3-pixel tolerance.
  - Captures contour quality and is more sensitive to thin-structure errors than region-based metrics.
4. Pixel-Share Analysis
  - Compare predicted vs ground truth pixel fractions for each class.
  - Helps assess whether the model systematically over- or under-segments ERM, forceps, or light tool, especially under extreme class imbalance.

### 3.8 Qualitative Evaluation

Quantitative metrics are complemented by visual inspection of segmentation outputs:

- “Good” cases where ERM and tools are clearly visible and predictions match annotations.
- “Challenging” cases with faint or unstained membranes, glare, or occlusions.

We construct panels showing the original image, annotation overlay, and model prediction overlay side by side for typical and difficult examples.



*Figure 6 Typical case: original frame with annotation and model prediction overlay*



*Figure 7 Challenging case: faint or partially unstained ERM with glare and tool overlap.*

## 4. Results

### 4.1 Quantitative results

On the curated test set, DeepLabv3+ at 512×512 resolution achieves the strongest overall performance:

- Present-only Dice (approximate):
  - ERM: ~0.77
  - Forceps: ~0.83
  - Light tool: ~0.76
- ERM Boundary-F1 @ 3 px: ~0.78
- Foreground mean mDice / mIoU: ~0.74 / 0.59
- ERM pixel share:
  - Prediction vs ground truth ≈ 0.159% vs 0.157%, indicating close alignment despite the membrane occupying less than 1% of pixels.

These numbers suggest that the model learns a meaningful representation of both membrane and instruments, capturing not just rough location but also approximate extent.

For UNet++ at 512×512, ERM Dice is slightly lower (around 0.70), with similar or somewhat lower performance on forceps and light tool. Foreground mean Dice/IoU is around 0.71 / 0.56, still acceptable but consistently below the DeepLabv3+ baseline.

At  $768 \times 768$  resolution, DeepLabv3+ underperforms the  $512 \times 512$  configuration, with lower Dice/IoU and weaker boundary quality for ERM. This indicates that simply increasing resolution does not guarantee better performance in the small-data, sparse-foreground regime and can actually make optimization more difficult.

## 4.2 Qualitative results

Visual inspection aligns with the quantitative findings:

- On well-stained and well-contrasted membranes, DeepLabv3+ produces smooth, contiguous ERM masks closely following the annotated boundaries. Instruments are also captured reliably, with good localization of forceps tips and light tool shafts.
- On faint or unstained membranes, predictions become more fragmented. The model may detect only portions of the membrane, particularly near the forceps where traction is apparent, while missing subtle peripheral regions.
- Instrument segmentation tends to be robust across lighting conditions, with occasional errors near strong specular highlights or at the edges of the field of view.

These observations underscore the importance of boundary-aware metrics and qualitative panels when evaluating models in thin-structure segmentation tasks.

## 5. Discussion

### 5.1 Feasibility of ERM Segmentation from Surgical Video

The primary question for this project is whether epiretinal membrane and instrument segmentation from intraoperative retinal video is feasible at all under realistic constraints: small dataset, thin low-contrast structures, and extreme foreground sparsity. The results indicate that the answer is yes—with caveats.

DeepLabv3+ at  $512 \times 512$  resolution achieves strong present-only Dice scores for ERM and instruments, and boundary-aware metrics suggest that contour alignment is often good on clearly visible membranes. The close match between predicted and ground truth ERM pixel share further suggests that the model has internalized a notion of how much membrane area typically appears when present.

At the same time, challenging cases remain problematic. Membranes that are barely visible even to human annotators, or that are partially obscured by instruments and glare, are often under segmented or missed. This is not surprising: even experienced clinicians encounter ambiguity in such frames, and the model cannot learn what is not reliably visible in the data.

### 5.2 Impact of Resolution and Data Scale

One somewhat counter-intuitive finding is that higher resolution ( $768 \times 768$ ) does not outperform  $512 \times 512$  on this pilot dataset. In fact, the higher-resolution configuration produces lower Dice and boundary metrics.

This can be interpreted in light of three interacting factors:

1. Dataset size: With only 300 frames in total and 48 test frames, increasing resolution does not add new semantic information; instead, it increases the number of pixels and degrees of freedom the model must optimize over.

2. Foreground sparsity: ERM and instruments together occupy less than 1% of the image. At higher resolutions, this proportion remains tiny, and the model must attend to very small islands of foreground amidst a large background.
3. Thin-structure optimization: Thin contours at high resolution are more sensitive to small misalignments, and the model may struggle to precisely delineate boundaries without sufficient data and regularization.

These observations reinforce an important lesson for medical image segmentation: more resolution is not always better, especially in small-data, highly imbalanced settings. Instead, careful selection of resolution, loss design, and metric choice can provide more reliable gains.

### 5.3 Metric Choice for Thin Structures

Traditional segmentation metrics such as mean Dice and IoU are informative but can obscure important details when foreground structures are thin and sparse. In this project, we found it useful to include:

- Present-only Dice, which avoids rewarding the model for correctly predicting “no ERM” on frames where the class is absent.
- Boundary-F1, which captures how well the model aligns with membrane contours.
- Pixel-share analysis, which tests whether the model is severely over- or under-segmenting classes relative to ground truth.

Together, these metrics paint a more nuanced picture: ERM segmentation is strong on clearly visible membranes, weaker on faint membranes, and relatively balanced in terms of area proportion. Such insights would be missed by overall pixel accuracy or unconditioned Dice alone

### 5.4 Role of Preprocessing and Annotation Support

A substantial portion of the effort in this project went into pipeline and annotation design, not just model training:

- SSIM-based de-duplication prevents over-representation of static frames and encourages models to learn from distinct surgical states.
- Blur and quality filtering ensures that extremely degraded frames do not dominate training.
- Annotator enhancement panels (CLAHE, edge maps, channel ratios) make faint ERM more visible, improving label consistency and quality.

These engineering choices are often under-reported in the literature but can significantly affect both annotation outcomes and model performance. In a small pilot setting, they are arguably as important as the choice of backbone.

## 6. Limitations

This work has several limitations that must be acknowledged:

1. **Data scale and diversity:** The dataset comprises only 300 frames from 7 surgeries, all from a single center. ERM appearance, staining protocols, and imaging conditions may differ at other sites. Results may not generalize without further data.

2. **Foreground sparsity and class imbalance:** ERM and instruments occupy far less than 1% of pixels. While composite losses and targeted metrics help, training remains challenging under such extreme imbalance.
3. **Annotation ambiguity:** Even with enhancement views, some ERM regions remain ambiguous. Small boundary disagreements between annotators and predicted masks can have a large effect on Dice and Boundary-F1.
4. **Lack of temporal modelling:** The models treat each frame independently and do not exploit temporal continuity. Temporal smoothing or sequence models could improve robustness but are beyond the scope of this pilot.
5. **Limited architectural exploration:** The study focuses on two convolutional backbones (DeepLabv3+ and UNet++). Transformer-based architectures, hybrid models, and more advanced loss functions remain to be explored.
6. **Non-public code and data:** Due to ongoing research and privacy constraints, source code and data cannot currently be shared, limiting external reproducibility. Within the VSI Lab, however, the pipeline is fully implemented and version-controlled.

## 7. Future Work

Within the VSI Lab and the broader retina-AI collaboration, this pilot benchmark provides a foundation for several lines of future work:

1. Expanded dataset
  - Incorporate additional ERM surgeries, including a broader range of staining protocols and imaging conditions.
  - Explore semi-automatic or active learning strategies to scale annotation efforts efficiently.
2. Improved membrane annotations
  - Perform iterative refinement of ERM boundaries with senior clinician review, especially at peel fronts and around forceps tips.
  - Consider labeling sub-regions (e.g., leading edge vs residual membrane) for richer supervision.
3. Advanced architectures
  - Evaluate transformer-based segmentation models and hybrid CNN-transformer architectures designed for medical images.
  - Investigate multi-scale context modules tailored to thin, elongated structures.
4. Boundary-aware objectives
  - Incorporate explicit boundary losses and uncertainty modeling at ERM edges.
  - Explore level-set-inspired losses or differentiable contour regularization.
5. Multi-task modeling
  - Extend the pipeline to jointly predict segmentation masks and surgical phases, leveraging shared encoders.
  - Investigate whether phase information improves segmentation quality (e.g., membranes during peeling vs completion).
6. Explainability and surgeon-facing interfaces
  - Develop saliency or attention visualizations focused on membrane regions.
  - Prototype overlays for potential use in training simulators or educational tools, subject to safety and validation.

## 8. Conclusion

This capstone project presents a pilot study of epiretinal membrane and instrument segmentation from intraoperative retinal surgery video. By curating a small but carefully designed dataset, implementing a comprehensive preprocessing and modeling pipeline, and evaluating modern segmentation architectures under realistic constraints, we demonstrate that:

- Pixel-level ERM and instrument segmentation from surgical video is technically feasible even with limited data.
- DeepLabv3+ at 512×512 resolution offers a strong baseline, with robust performance on clearly visible membranes and instruments.
- Thin-structure metrics and qualitative inspection are crucial to understanding behavior on challenging cases.
- Preprocessing, deduplication, annotation support, and surgery-level splitting are critical components of a trustworthy benchmark, not merely implementation details.

While far from clinical deployment, the work establishes a benchmark-style framework for intraoperative retinal surgery segmentation in an underrepresented domain. Within the VSI Lab, this framework can be extended to larger datasets, advanced models, boundary-aware objectives, and multi-task formulations that jointly model segmentation and surgical workflow.

## Acknowledgements

This project was conducted in the Vision, Systems & Intelligence (VSI) Lab at the University of Arizona as part of the M.S. Information Science capstone.

The author thanks Dr. Eungjoo Lee for guidance on project design, model selection, and evaluation; Dr. Kozak and clinical collaborators for domain expertise, feedback on ERM identification, and clarifying what counts as membrane during peeling; and fellow VSI Lab members for annotation support, pipeline feedback, and helpful discussions throughout the project

## References

1. Twinanda, A. P. et al. "EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.
2. Allan, M. et al. "2017 Robotic Instrument Segmentation Challenge." 2019.
3. Bodenstedt, S. et al. "Comparative Evaluation of Instrument Segmentation Methods in Minimally Invasive Surgery." *IEEE Transactions on Medical Imaging*, 2018.
4. Chinedu et al. "CholecTrack20: A Dataset for Multi-Class Multiple Tool Tracking in Laparoscopic Surgery." 2023.
5. Rogeria et al. "Feature Tracking and Segmentation in Real Time via Deep Learning in Vitreoretinal Surgery: A Platform for Artificial Intelligence-Mediated Surgical Guidance." 2022.
6. Ming et al. "OphNet: A Large-Scale Video Benchmark for Ophthalmic Surgical Workflow Understanding." 2024.
7. Ayhan (Murat) et al. "Interpretable Detection of Epiretinal Membrane from Optical Coherence Tomography with Deep Neural Networks." *Scientific Reports*, 2024.
8. David et al. "The Role of Artificial Intelligence in Epiretinal Membrane Care: A Scoping Review." *Ophthalmology Science*, 2025.
9. Carlà et al. "Smartphone Augmented En-Face Guided Epiretinal Membrane Peeling: A 3D Ngenuity Tool for Customized Treatment." *Retina*, 2025.
10. Run Zhou et al. "Intraoperative Augmented Reality for Vitreoretinal Surgery Using Edge Computing." 2025.
11. Onur et al. "Automated Detection of Retinal Detachment Using Deep Learning-Based Segmentation on Ocular Ultrasonography Images." 2025.
12. Charumathi et al. "A Deep Learning Algorithm to Detect Chronic Kidney Disease from Retinal Photographs in Community-Based Populations." 2020.
13. Guanrong et al. "Association of Retinal Age Gap with Chronic Kidney Disease and Subsequent Cardiovascular Disease Sequelae: A Cross-Sectional and Longitudinal Study from the UK Biobank." 2024.
14. Bjorn et al. "Deep Learning Algorithms to Detect Diabetic Kidney Disease from Retinal Photographs in Multiethnic Populations with Diabetes." 2023.
15. Youngmin et al. "Diagnosis of Chronic Kidney Disease Using Retinal Imaging and Urine Dipstick Data: Multimodal Deep Learning Approach." 2025.
16. Yuhe et al. "Performance of Deep Learning for Detection of Chronic Kidney Disease from Retinal Fundus Photographs: A Systematic Review and Meta-Analysis." 2024.