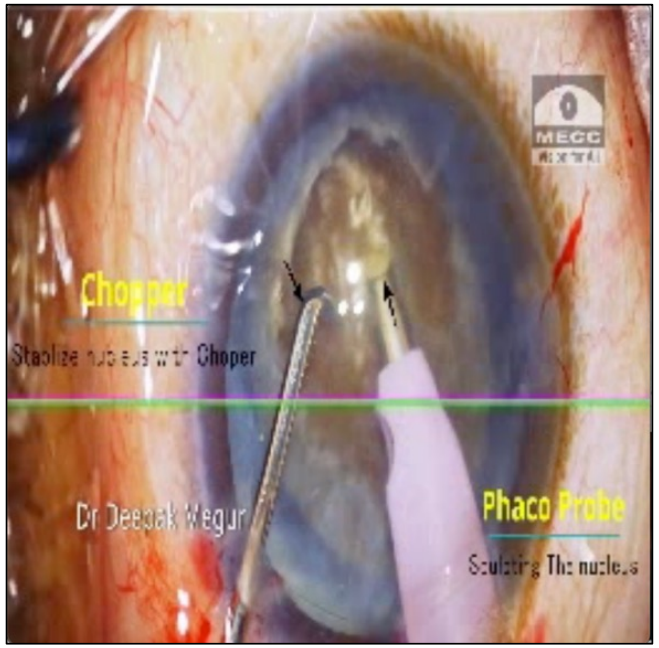
APTOS 2025 Big Data Competition
- Phase Recognition of Surgical Videos Using ML
1st Place Solution : Segmentation-Informed Ensemble

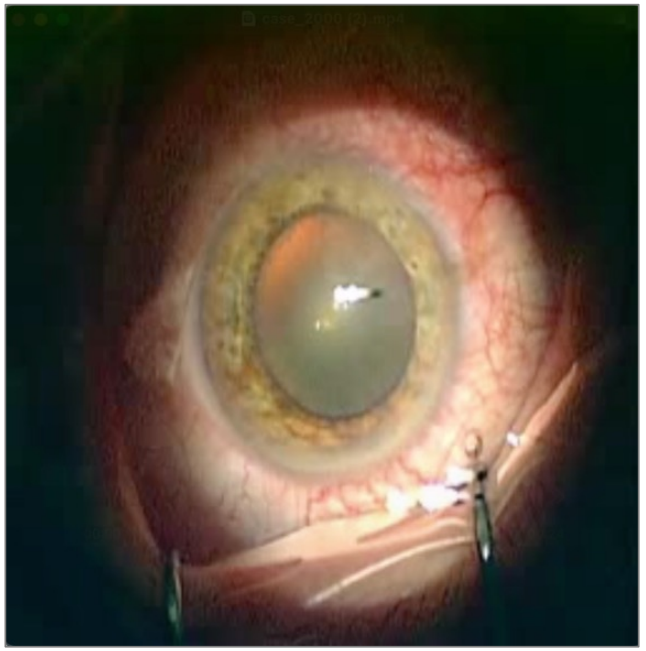
Hiroaki Fukuse, MD
Resident (PGY-2)
Center for Medical Education and Training
Kobe University Hospital, Japan

Comparison of Video Features

Train Dataset		Test Dataset
	Videos	
496 videos		100 videos
	Frame Rate	
about 30 fps		1 fps
	Quality	
Noisy		Clean
	Source	
YouTube[1]		Unknown



case_0024.mp4



case_2000.mp4

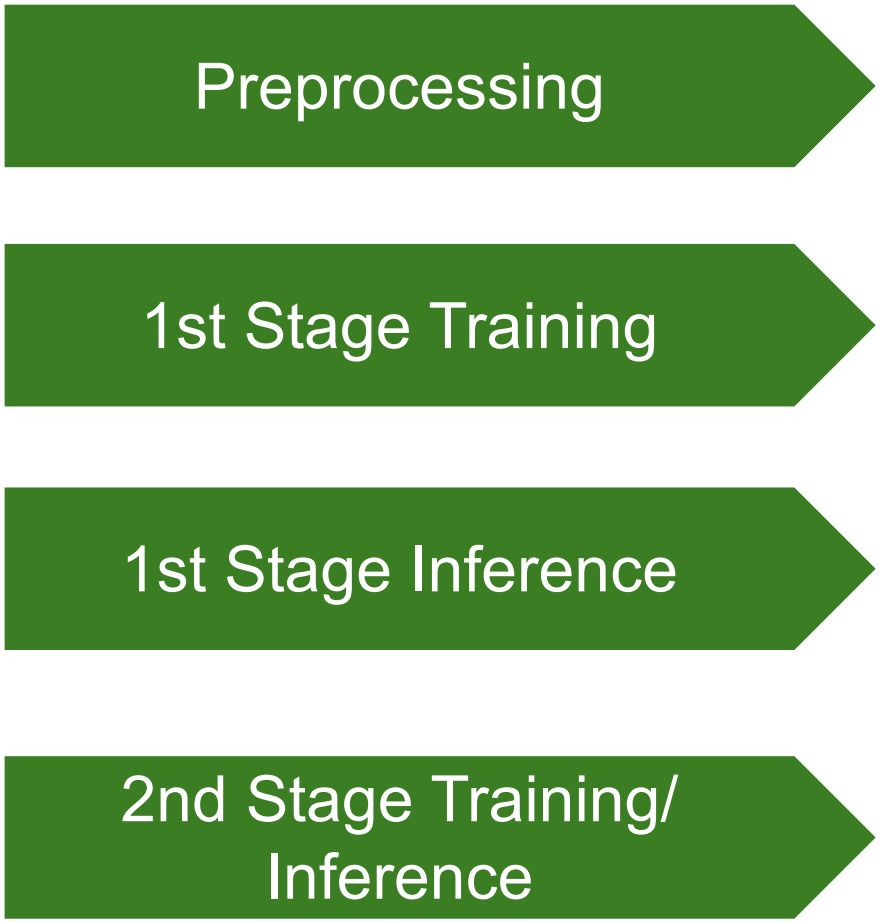
[1]OphNet: A Large-Scale Video Benchmark for Ophthalmic Surgical Workflow Understanding

Video Features：Insights

Binary Classification
Can Train vs Test be separated?

- Different FPS between Train and Test
  →At 1 fps, **CNN-RNN > 3D CNN ?**
- Iris–pupil complex & tool tips are small but crucial
  → **Avoid aggressive down-scaling ?**
- Train noisy, Test clean
  → **Strong augmentation may be less critical ?**
- 35 classes classification
  → High variance; **increasing folds** improve stability ?
- Label distribution shifts along the surgical timeline
  → **Time encoding might help ?**

Pipeline for Deep Learning Model Development

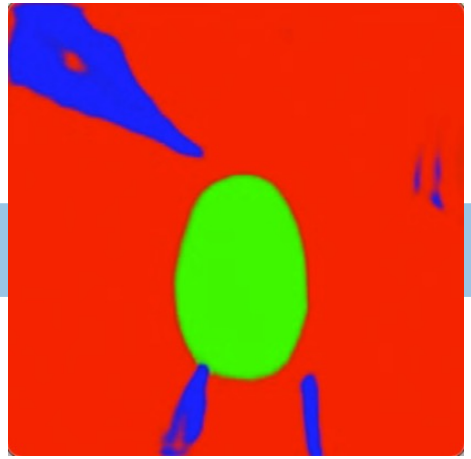


Pipeline for Deep Learning Model Development

Preprocessing → 1st Stage Training → 1st Stage Inference → 2nd Stage Training/ Inference

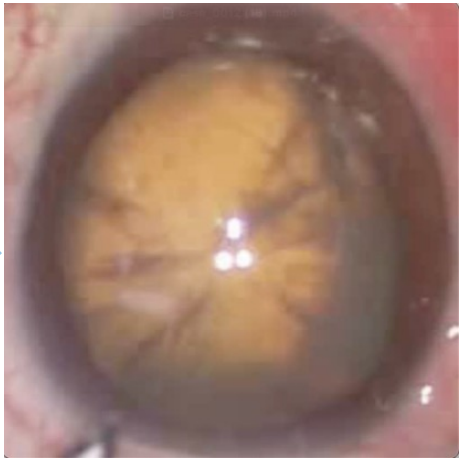
Raw video

U-Net *



Segmentation

Crop



Cropped video

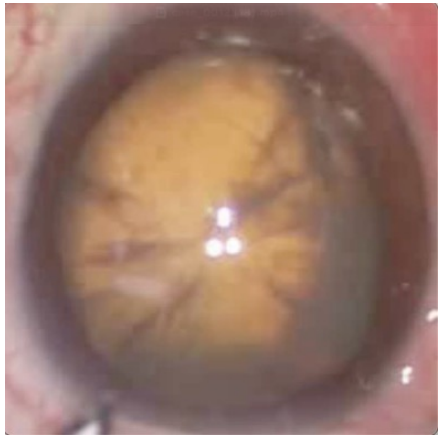
* Approximately 200 frames were randomly sampled from the training videos and manually segmented for the iris–pupil complex and surgical instruments. A U-Net model was trained on these annotations and then used to infer segmentations on both the training and test videos.

Preprocessing | **1st Stage Training** | 1st Stage Inference | 2nd Stage Tr… Inferenc…

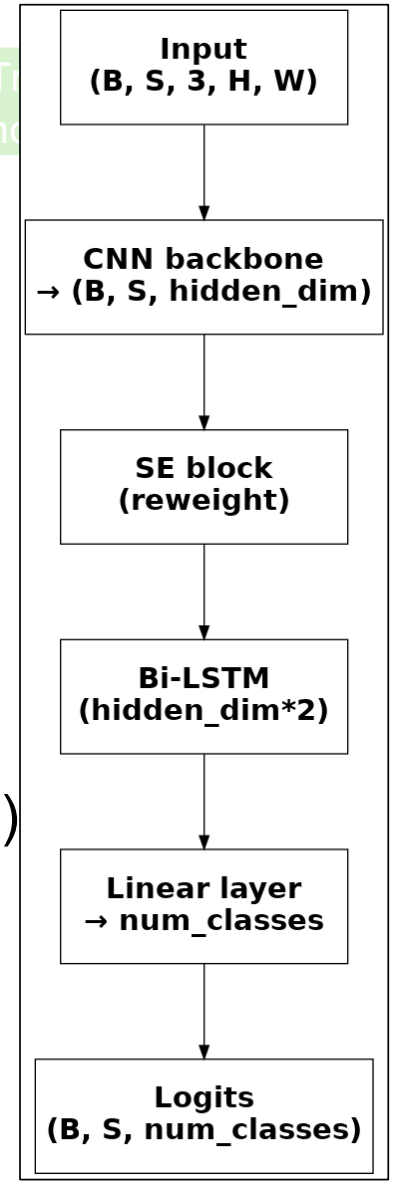
Model①

- Input :
  Raw image frames (BS,10,3,**512,512**)
- CNN backbone:
  tf_efficientnetv2_m.in21k_ft_in1k
- SE block
- Transition loss as an auxiliary loss
- Stratified Group 10-fold

Model②



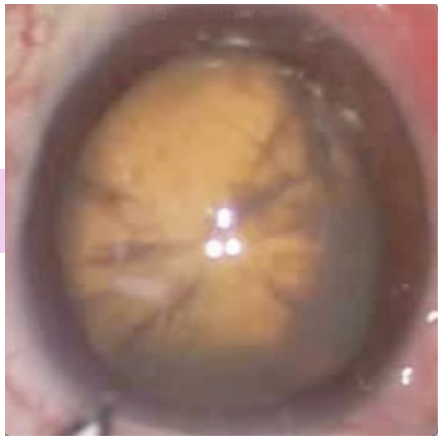
- Input :
  Cropped image frames (BS,10,3,**256,256**)
- CNN backbone:
  convnext_base.fb_in22k_ft_in1k_384
- No SE block
- Transition loss as an auxiliary loss
- Stratified Group 10-fold

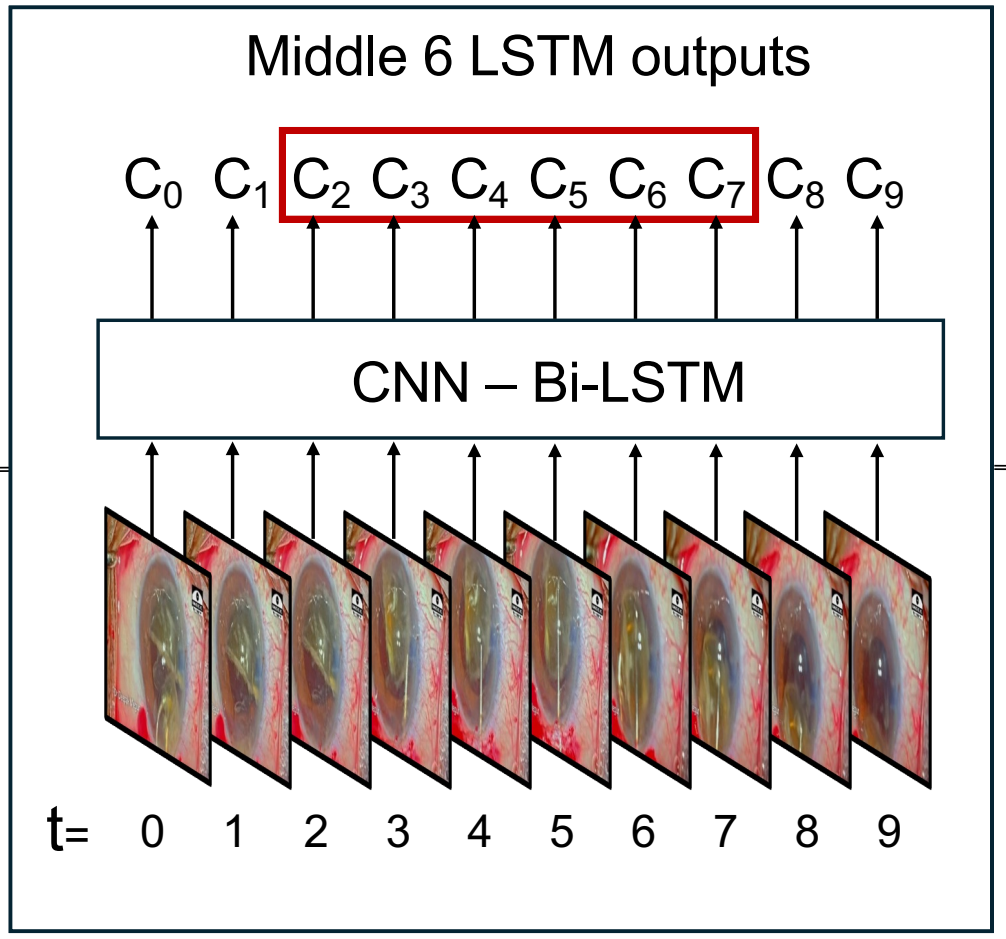


Preprocessing | 1st Stage Training | **1st Stage Inference** | 2nd Stage Training/ Inference

Model①

Model②



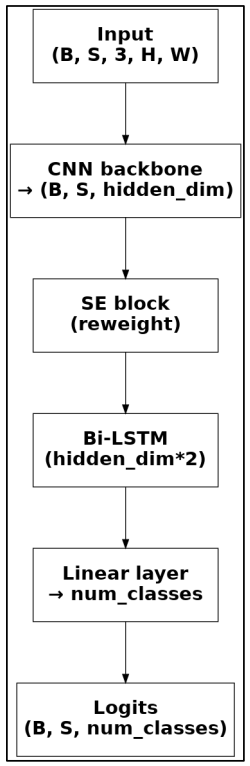


LB: 0.5862

LB: 0.5692

① + ② → 0.6213

Preprocessing | 1st Stage Training | 1st Stage Inference | 2nd Stage Training/ Inference

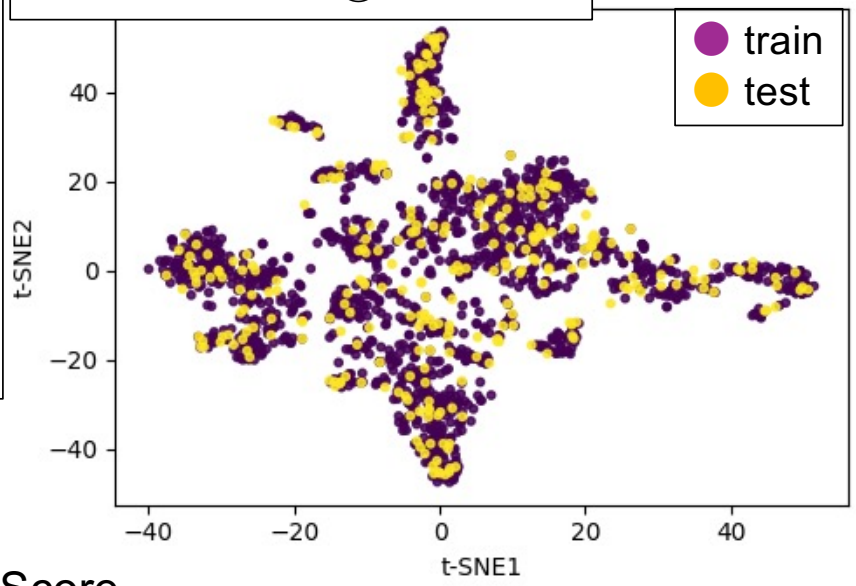


Model③：with Pseudo-Label
tf_efficientnetv2_m.in21k_ft_in1k

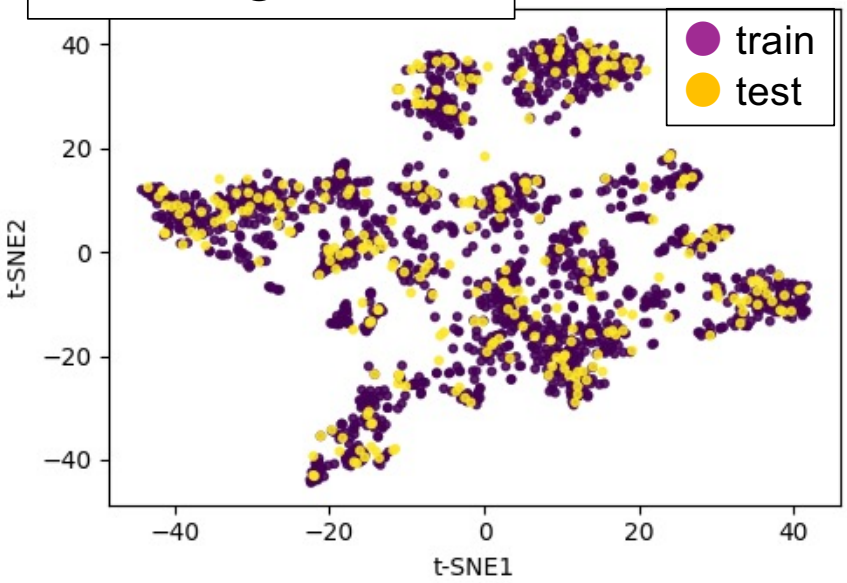
Model④：with Pseudo-Label
tf_efficientnet_b5.ns_jft_in1k

LB: 0.6384
LB: 0.6245
→ LB: 0.6394

Non-PL Model ① Features



PL Model③ Features



Silhouette Score
(●Train vs ●Test)

-0.008 > -0.021

Techniques that worked for me

My Baseline Model LB **0.5368**

CNN-Bi-LSTM with Transition loss & SE block

+ Middle 6 outputs	LB: **0.5509**	+ 0.0141	
+ 5fold → 10fold	LB: **0.5862**	**+ 0.0353**	Model①
+ 1st stage ensemble	LB: **0.6213**	**+ 0.0351**	Model①+②
+ Pseudo-Label	LB: **0.6384**	+ 0.0171	Model③
+ 2nd stage ensemble	LB: **0.6394**	**(Final Score)**	Model③+④

Video Level Accuracy 0.7889

Techniques that did Not work for me

- Video MAE
- Time encoding
- Larger chunks
- Image augmentation
- Replaced LSTM with RNN or GRU
- Used more complex LSTM architectures

(e.g., multiple layers, residual connections)

- CycleGAN (convert train data into test data)

. . .

and other techniques

Conclusion

- The CNN–LSTM pipeline was effective for processing low-frame-rate (1 fps) surgical videos.
- Cropped videos focusing on the iris–pupil complex, aided by U-Net based segmentation, were especially useful in ensemble models.
- Explicitly guiding the model to focus on important regions may be crucial, especially under dataset shift conditions.

Q&A / Contact

Hiroaki Fukuse, MD Kobe University

🔗 GitHub: hfukuse/aptos2025_1st_place_solution

Email: hif_med-ml@yahoo.co.jp


Thank you for your attention.