Neurosurgical Phase Recognition for Aneurysm Clipping via Fine-Tuned CNN and Temporal Models

Emily Oberleitner Veronica Wang Nicole Wong

Department of Computer Science, Stanford University

{eober, vwang2, nicolew6}@stanford.edu

Motivation & Core Challenges

- **Intelligent Operating Rooms:** Automatic surgical phase recognition is the foundation for context-aware intraoperative assistance, safety auditing, and efficient video summarization for residency training pipelines.
- **Extreme Visual Homogeneity:** Unlike general surgery showing highly contrasting multi-organ views, neurosurgical surgeries present as a continuous, uniform landscape of tissue with fine movement, stripping standard models of distinct anatomical landmarks or actions.
- **Intraoperative Artifacts:** Real-world video is plagued by severe visual noise such as rapid microscope focus/position adjustments, periodic saline irrigation, lens blood, and debris.
- **Long-Range Temporal Drift:** Traditional networks relying on RNNs to parse video sequences may suffer from severe memory decay over multi-hour procedures and lack the self-attention mechanisms needed to dynamically weight non-adjacent frame relationships.

To resolve these limitations, we adapt an **Attention-Regularized Transformer (OperA)** framework tailored specifically to the harsh visual environment of neurosurgical videos:

- **Decoupled Architecture:** We utilize a ResNet-50 backbone for spatial feature extraction, passing frame embeddings into a 2-layer causally masked transformer network to capture global, long-range sequential context without future information leakage.
- **Attention Regularization (L_{reg}):** To handle intraoperative noise, we implement a specialized regularization loss. It penalizes the transformer if it assigns high attention weights to frames where the underlying spatial backbone exhibits low confidence or high prediction error.

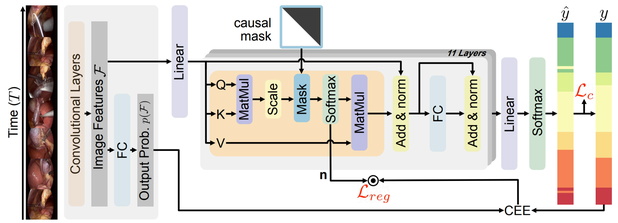

Figure 1. OperA model architecture.

Background & Related Works

- **Early Paradigms:** Initial approaches utilized Hidden Markov Models (HMMs) and basic ANNs, eventually shifting toward deep feature extraction with *EndoNet*, which first enabled simultaneous recognition of phases and tool presence.
- **The Spatial-Temporal Shift:** Whilst frame-wise CNNs learn critical visual features, they lack global sequence context. To bridge this, architectures like *EndoLSTM*, *SV-RCNet*, and *TeCNO* (Multi-stage TCNs) introduced temporal constraints, significantly boosting F1 scores by modeling neighboring frame relationships.
- **Phase-Defining Focus:** *OperA* uses attention regularization to prioritize "high-information" frames, a technique we adapt to counter the visual homogeneity of neurosurgery.

Dataset & Features

- **Data Core:** 48 aneurysm neurosurgery videos manually annotated in CVAT at 1 fps, generating 41,604 frames.
- **Surgical Milestones:** 4 explicit, sequential phases: (1) *Brain Exposure* (18.1%), (2) *Parent Vessel Identification* (25.1%), (3) *Neck/Dome Identification* (20.1%), and (4) *Clipping* (36.8%).
- **Feature Splits:** Training is done with 64-frame sliding windows and validation is evaluated using strict video-level data. We use a 80-20 train/val ratio.

Methodology

Causal Mask: To implement data leakage and simulate real surgical environments, we implement a causal mask where

$$\text{Mask}_{ij} = \begin{cases} 0 & \text{if } i \geq j \\ -\infty & \text{otherwise} \end{cases}$$

Attention Regularization Loss (L):

$$L = L_{CE} + \lambda L_{reg}$$

Where L_{CE} represents cross-entropy loss with inverse-frequency class weights ($w_c \propto 1/N_c$) and label smoothing ($\epsilon = 0.1$). L_{reg} explicitly penalizes the self-attention network if it assigns high attention weights to frames where the spatial backbone exhibits low confidence or high prediction error.

Viterbi Decoupled Post-Processing: Finds the structurally optimal sequence under a first-order Markov model:

$$Y^* = \arg\max_Y \prod_{t=1}^{T} P(v_t|y_t) \cdot P(y_t|y_{t-1})$$

We employ a tolerance window of 5 seconds (5 frames) around phase boundaries, excluding frames within this region from accuracy computation to avoid penalizing predictions near inherently ambiguous transitions.

Architecture Hyperparameters

Component	Configuration
Backbone	ResNet-50, fine-tuned end-to-end
Transformer	11 layers, 4 heads, d=256, p=0.5
MS-TCN	2 stages × 8 dilated layers, 128 filters, p=0.5
Masking	Causal (no future leakage)
Post-processing	Viterbi decoding (learned transition matrix)

Training Protocol Settings

- **Optimization Engine:** AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-3}).
- **Learning Rate Profile:** Linear warmup across 5 epochs followed by a cosine annealing learning rate schedule dropping to $\eta_{min} = 10^{-6}$ with a maximum peak $\eta_{max} = 5 \times 10^{-4}$.
- **Fine Tuning:** We evaluated two feature regimes: (1) frozen ImageNet ResNet-50 features and (2) fine-tuned ResNet-50 features trained on labeled surgical frames. t-SNE visualization of frozen features showed no meaningful cluster structure across the four surgical phases, motivating fine-tuning.

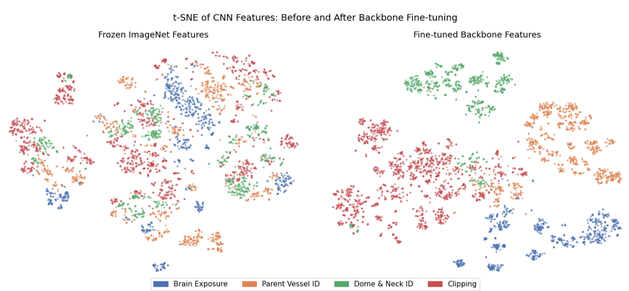

Figure 2. t-SNE of frozen ImageNet features (left) vs fine-tuned backbone features (right) from 4,000 randomly sampled frames.

Results & Discussion

Metrics are reported at the best validation loss epoch. Edit and F1 use Viterbi post-processing where applicable.

Model Structure	Features	Acc	Edit	F1@10	F1@50
CNN Baseline	Frozen	0.956	0.000	0.000	0.000
LSTM	Frozen	0.116	0.024	0.018	0.004
Transformer	Frozen	0.533	0.015	0.016	0.004
MS-TCN + Viterbi	Frozen	0.525	0.467	0.553	0.340
MS-TCN + Vit. (learned)	Frozen	0.503	0.583	0.585	0.390
MS-TCN + Viterbi	Trainable	**0.957**	0.680	0.809	0.714
Transformer + Vit. (OperA)	Trainable	0.936	**0.810**	**0.895**	**0.789**

Table 1. Performance metrics across baseline and proposed architectures.

- **The Baseline Paradox:** The frozen CNN achieves high frame accuracy (95.6%) but 0.000 Segmental F1@10 from overguessing the majority class on highly imbalanced datasets.
- **Fine-Tuning Impact:** Fully training the ResNet-50 visual backbone on neurosurgical frames was the most powerful driver of accuracy, increasing the OperA Transformer's F1@10 score from 0.533 to 0.895.
- **Remaining errors:** Errors concentrate near the boundaries between *Neck/Dome Identification* and *Clipping*

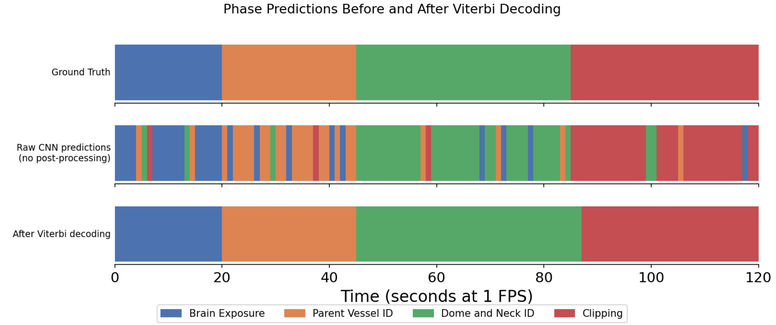

Figure 3. Phase timeline for a 120-second video clip. Ground truth (top) shows the ideal phase sequence. Raw CNN predictions (middle) exhibit temporal flickering. After Viterbi decoding (bottom), the sequence is constrained to forward-only transitions, following ground truth.

Conclusion & Future Work

- **Best Performance:** The causal transformer paired with a learned Viterbi transition matrix achieved the best overall sequence performance (F1@10 = 0.944, F1@50 = 0.789), competitive with SOTA systems trained on much larger surgical datasets.
- **Algorithmic Refinements:** Explore boundary-aware loss functions to penalize misaligned segment endpoints at training time and incorporate explicit tool detection.
- **Clinical Deployment:** Evaluate real-time, online inference latency at 1 FPS on physical clinical hardware to transition the causally masked transformer framework into the operating room.

References & Acknowledgments

Twinanda et al. "EndoNet: A deep architecture for recognition of anatomical structures and tools." *IEEE TMI*, 2016.
Czempiel et al. "TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks." *MICCAI*, 2020.
Czempiel et al. "OperA: Attention-regularized transformers for surgical phase recognition." *MICCAI*, 2021.

Data Contribution: We thank Dr. Jinendra Ekanayake for providing the neurosurgical dataset and expert labels.