Boundary-Aware FACT: Explicit Boundary Supervision for Frame–Action Cross-Attention

SICS-155 Surgical Phase Recognition Challenge - Team AtlasVision

Yasser El Jarida   Youssef Iraqi   Loubna Mekouar

2025

College of Computing, University Mohammed VI Polytechnic
Benguerir, Morocco

SICS-155 Challenge Overview

Challenge Details

SICS-155: 155 videos, 19 phases, 100 train / 15 test

Key Challenge: Temporal boundary ambiguity at phase transitions

- Over-segmentation: short spurious segments
- Boundary blurring between adjacent actions
- Need for stable phase segmentation

Team AtlasVision Approach

Strategy: Boundary-aware training for better temporal consistency

Base Model: FACT with I3D features

Innovation: Auxiliary boundary head for transition prediction

MICCAI2025 Daejeon REPUBLIC OF KOREA

Figure 1: MICCAI 2025

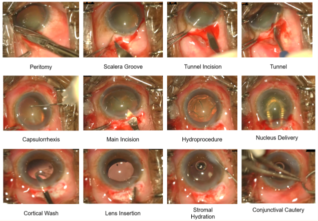


Figure 2: SICS phases

Previous Methods & Validation Results

MS-TCN++ with custom features

VideoMAE-v2 features:

- Pretrain: Cataract-1K + OphNet (2024), then finetune on Cataract101
- Segmenter: MS-TCN++ (MSTCN2-style temporal convs)
- Outcome: lower Acc/F1/Edit; unstable early training

I3D features + MS-TCN++: improved but $<$ **80% Acc**

Surgformer (HTA head)

- Microscopic + macroscopic temporal attention for long-range dependencies
- **Acc 82%** on validation, but **poor F1/Edit**

Motivation for FACT

Combine convolutional efficiency with transformer long-range modeling via cross-attention

Notes

- VideoMAE-v2 + MS-TCN++ unstable from early epochs
- I3D + MS-TCN++ improved stability but $<$ 80% Acc
- Surgformer: long-range modeling, but weak F1/Edit

Approach & FACT Baseline

High-level approach

- Backbone features: I3D
- Temporal model: FACT (frame CNN branch + action-token transformer)
- Information exchange: bidirectional cross-attention between branches
- Inference: merge token-derived posteriors with frame logits via a learned weight

How it works (simple view)

- **Frame branch (convolutions):** captures local motion/appearance and produces per-frame class scores efficiently.
- **Action branch (tokens + transformer):** a *small set of learnable action tokens* model long-range structure and segment-level context.
- **Cross-attention (both directions):** tokens attend to frames to align with segments; frames attend to tokens to receive high-level guidance.
- **Final prediction:** combine guidance from tokens with the frame branch for stable, accurate framewise labels.

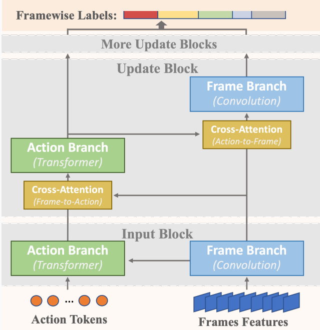


Figure 3: FACT baseline

Boundary-aware Extension (Architecture & Loss)

Loss formulation

Boundary detection (BCE):

$$\mathcal{L}_{\text{BCE}} = \frac{1}{T} \sum_{t} \left(- y_b(t) \log p_b(t) - (1 - y_b(t)) \log(1 - p_b(t)) \right)$$

Boundary-weighted TV: $\mathcal{L}_{\text{TV}}^{\text{w}} = \dfrac{1}{T-1} \displaystyle\sum_{t=1}^{T-1} (1 - p_b(t))^{\gamma} \, \| \log \mathbf{p}_{t+1} - \log \mathbf{p}_t \|_2^2$

Combined per-block: $\mathcal{L}_{\text{block}} = \mathcal{L}_{\text{frame}} + \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{attn}} + \alpha \, \mathcal{L}_{\text{TV}}^{\text{w}} + \beta \, \mathcal{L}_{\text{BCE}}$

Intuition

- Stabilize interiors; allow sharp changes at true transitions
- Gate smoothing by $(1 - p_b(t))^{\gamma}$
- No inference cost or parameter changes

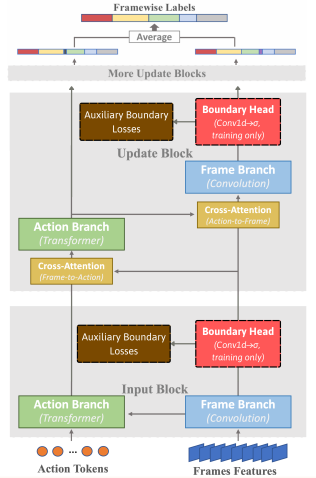


Figure 4: Boundary-aware head placement

Experimental Setup

Dataset & Features

SICS-155: 100 train / 15 test videos, 19 phases

- 960×540 resolution @ 30 FPS
- I3D features: 1024-D spatiotemporal embeddings
- Extracted at stride $sr = 3$

Training Strategy

Warm Start: Vanilla FACT baseline

- Load shared weights, initialize boundary heads
- Learning rate: 1×10^{-4}
- Merge weight: $w = 0.50$

Hyperparameter Search: W&B sweeps

- Boundary loss weight: $\{1.0, 1.5, 2.0\}$
- TV exponent: $\{2.0, 3.0\}$
- Smoothing weight: $\{1.0, 2.5, 5.0\}$

Hardware

- Single NVIDIA RTX 6000 Ada Generation GPU
- Efficient training with warm initialization

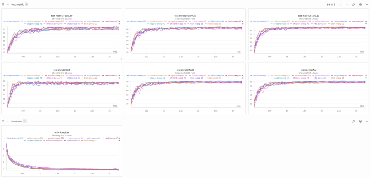
Hyperparameter Tuning (W&B Sweeps)

Search spaces

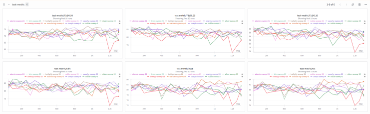
- β (boundary BCE): $\{1.0, 1.5, 2.0\}$
- γ (TV exponent): $\{2.0, 3.0\}$
- α (smoothing): $\{1.0, 2.5, 5.0\}$
- Frame feature stride sr: $\{1, 3, 5\}$
- Merge weight w: fixed 0.50; LR 1×10^{-4}; $M = 36$

Selected configuration

$\beta = 1.0,\ \gamma = 2.0,\ \alpha = 5.0,\ sr = 3,\ w = 0.50,\ M = 36$



w/o boundary



with boundary (warm start)

Results & Performance (Validation & Test)

Test Set Results

SICS-155 Challenge Submission:

- **Accuracy: 82%** (Rank #2 on leaderboard)
- Consistent performance across test videos
- Boundary-aware variant showed improvements

Validation Set (Public)

Method	Acc (%)	F1@0.50	Edit
FACT	82.8	77.1	**86.9**
+ boundary	**84.1**	**78.6**	86.3

Qualitative Improvements

- Cleaner phase transitions
- Fewer short spurious segments
- Better boundary adherence

Key Insights

- Boundary awareness helps temporal consistency
- Minimal computational overhead
- Warm start crucial for stability

Qualitative Results

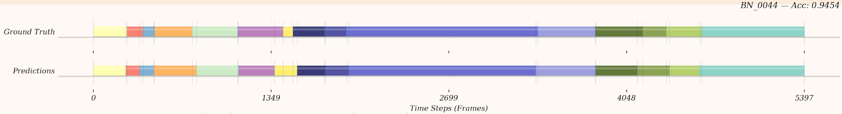


Figure 5: Ground truth (top) vs. predictions (bottom) showing cleaner boundaries and reduced over-segmentation

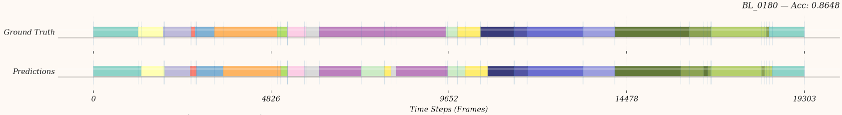


Figure 6: Additional sequence showing consistent boundary alignment

Ablation Study & Analysis

Component Analysis

Boundary Head Impact:

- BCE loss alone: marginal improvement
- Weighted TV loss alone: moderate improvement
- Combined: best performance (+1.3% accuracy)

Training Dynamics:

- Warm start critical for stable convergence
- Learning rate 1×10^{-4} optimal
- Higher rates cause training instability

Hyperparameter Sensitivity

- $\gamma = 2.0$ optimal for TV exponent
- $\beta = 1.0$ best boundary loss weight
- Merge weight $w = 0.50$ most stable

Computational Cost

Training:

- +1 Conv1D per block
- Negligible parameter increase

Inference:

- Identical to baseline
- No runtime overhead

Limitations

- Single dataset (SICS-155)
- I3D features only
- Binary boundary supervision

Conclusion & Future Work

Key Contributions

Methodological:

- Strategic boundary head integration into FACT
- Boundary-weighted temporal smoothing loss
- Training-only overhead design

Results:

- 82% accuracy on test set (Rank #2)
- Qualitative reduction in over-segmentation
- Consistent gains with proper warm start

Clinical Impact

- Better phase boundary detection for surgical training
- Improved quality assurance tools
- Cost-effective for resource-limited settings

Future Work

- Alternative boundary integration strategies
- Cross-dataset generalization
- Real-time deployment optimization
- Clinical validation trials

Boundary-aware training improves temporal consistency at no computational cost.

Thank You

Yasser El Jarida
UM6P College of Computing
yasser.eljarida@um6p.ma

Youssef Iraqi
UM6P College of Computing
youssef.iraqi@um6p.ma

Loubna Mekouar
UM6P College of Computing
loubna.mekouar@um6p.ma

Slides: https://yasser.sh/talks/miccai_sics155/