Boundary-Aware FACT: Explicit Boundary Supervision for Frame–Action Cross-Attention on SICS-155

Yasser El Jarida, Youssef Iraqi, and Loubna Mekouar

College of Computing, University Mohammed VI Polytechnic
Benguerir, Morocco
{yasser.eljarida,youssef.iraqi,loubna.mekouar}@um6p.ma

Abstract. This paper presents a Boundary-Aware extension of FACT (Frame–Action Cross-attention Temporal modeling) for surgical phase segmentation on SICS-155. FACT couples a frame branch (temporal convolutions) with a compact set of learnable action tokens (transformer), linked through cross-attention. We retain the original design and add a lightweight boundary head (Conv1D→Sigmoid) alongside each frame branch. The head is used only during training: it predicts per-frame boundary probabilities to supervise a boundary BCE term and to reweight the temporal smoothing term into a boundary-aware total-variation loss, promoting smooth interiors while permitting sharp changes at true transitions. Inference is unchanged. On the SICS-155 split (100 train / 15 test; 19 phases), a tuned FACT baseline reaches 82.8% frame accuracy; the boundary-aware variant improves to 84.1% and qualitatively reduces over-segmentation. We also report negative trials with a VideoMAE-v2 feature pipeline (self-supervised pretraining on Cataract-1K and OphNet, fine-tuning on Cataract101, then MS-TCN++), which underperformed I3D features in this setting. The proposed regularization injects transition awareness at negligible cost while keeping test-time computation intact.

Keywords: Action segmentation · Cross-attention · Temporal modeling · Surgical workflow · Boundary detection

1 Introduction

Accurate framewise action segmentation remains difficult at phase transitions, where models often fragment labels into short spurious segments or blur boundaries between adjacent actions. In surgical video, precise phase boundaries are central to workflow understanding and downstream assistance.

Two families of approaches dominate. Convolutional temporal models, typified by MS-TCN and MS-TCN++, expand temporal receptive fields with dilated 1D convolutions and offer a strong inductive bias at modest cost, performing competitively on SICS-style datasets [3–5]. Transformer-based models

target long-range dependencies with attention; Surgformer emphasizes hierarchical temporal attention [7]. FACT lies between these extremes: a temporal-convolutional frame branch interacts with a compact set of learnable action tokens processed by a lightweight transformer, and bidirectional cross-attention mediates information exchange between tokens and frames [1]. Boundary-aware training, as in ASRF, addresses transition quality directly by predicting boundary probabilities and using them to regularize temporal smoothing [2].

This work retains FACT's cross-attentional architecture and injects boundary awareness only during training. A lightweight boundary head runs in parallel to each frame branch to predict per-frame transition probabilities that supervise a boundary loss and gate the temporal smoothing term, encouraging stability within segments while permitting sharp changes at true transitions. The inference pipeline is unchanged, with no extra parameters, memory, or runtime. On SICS-155, this boundary-aware variant improves accuracy and qualitatively reduces over-segmentation. Our contributions are a strategically placed integration of boundary supervision into FACT, a loss formulation that couples boundary prediction with boundary-aware smoothing, and an empirical study on SICS-155 demonstrating consistent gains.

2 Materials and Methods

Experiments follow the SICS-155 split (100 train / 15 test; 19 classes). Inputs are 1024-D I3D features extracted at stride $\mathtt{sr} = 3$. We also evaluated a VideoMAE-v2 pipeline that used self-supervised pretraining on Cataract-1K and OphNet, followed by fine-tuning on Cataract101 and segmentation with MS-TCN++ [8, 4]. This alternative produced lower accuracy and edit scores with unstable early training, so all main results use I3D features.

We briefly recall the FACT formulation. Let $\mathbf{X} \in \mathbb{R}^{T \times D}$ denote frame features and $\{\mathbf{a}_m\}_{m=1}^{M}$ the learnable action tokens of dimension H. Each block comprises a frame branch (MSTCN2-style temporal convolutions [3, 4]), an action branch (a transformer operating on tokens), and cross-attention maps between frames and tokens. The final block produces per-frame logits $\mathbf{z}_t \in \mathbb{R}^C$ and token logits $\mathbf{u}_m \in \mathbb{R}^{C+1}$ (including a null class). Per-frame posteriors combine the token-to-frame projection with frame-branch predictions,

$$\mathbf{p}_t = w\, \mathbf{p}_t^{\text{tok}\to\text{frm}} + (1 - w)\, \text{softmax}(\mathbf{z}_t), \quad \hat{y}_t = \arg\max_c \mathbf{p}_{t,c}, \tag{1}$$

where $\mathbf{p}_t^{\text{tok}\to\text{frm}} \in \mathbb{R}^C$ and $\text{softmax}(\mathbf{z}_t) \in \mathbb{R}^C$ are class posteriors per frame and $w \in [0, 1]$ weights the token-to-frame term as in FACT. Baseline training uses frame cross-entropy with background weighting, token cross-entropy with Hungarian matching based on a cost that mixes token class probability and attention IoU, attention alignment that concentrates mass on matched segments, and temporal smoothing via squared differences of log-posteriors across adjacent frames.

We extend the frame branch by adding a parallel Conv1D $\times 3$ layer (with padding 1) followed by a sigmoid activation, yielding a per-frame boundary

probability $p_b(t) \in [0,1]$ using the same features as the classifier. For supervision, boundary labels are derived from frame annotations such that $y_b(t) = \mathbb{1}\{\, y_t \neq y_{t-1} \,\}$. This setup introduces two additional loss terms used only during training. The first is a binary cross-entropy loss for boundary detection,

$$\mathcal{L}_{\text{BCE}} = \tfrac{1}{T} \sum_t \Big(- y_b(t) \log p_b(t) - (1 - y_b(t)) \log \big(1 - p_b(t)\big) \Big). \tag{2}$$

Second, a boundary-weighted total variation that replaces uniform smoothing,

$$\mathcal{L}_{\text{TV}}^{\text{w}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \big(1 - p_b(t)\big)^{\gamma} \, \|\log \mathbf{p}_{t+1} - \log \mathbf{p}_t\|_2^2 \,, \tag{3}$$

where $\mathbf{p}_t \in (0,1)^C$ is the class posterior vector at time t and $\log \mathbf{p}_t \in \mathbb{R}^C$; the exponent $\gamma > 0$ controls how strongly boundaries relax the penalty. The per-block objective becomes

$$\mathcal{L}_{\text{block}} = \mathcal{L}_{\text{frame}} + \mathcal{L}_{\text{token}} + \mathcal{L}_{\text{attn}} + \alpha \, \mathcal{L}_{\text{TV}}^{\text{w}} + \beta \, \mathcal{L}_{\text{BCE}}, \tag{4}$$

and losses are averaged across blocks. The boundary head is not used at inference, and decoding follows Eq. (1).

Figure 1 summarizes the complete data flow and the placement of the boundary head. The head taps the frame-branch features immediately before the per-class linear layer and outputs a scalar boundary probability $p_b(t)$. During training, this signal supervises boundary detection (BCE) and gates the temporal smoothing term in Eq. (3) via $(1 - p_b(t))^{\gamma}$, which encourages stability inside segments while relaxing the penalty at true transitions. At inference the head is ignored and decoding follows Eq. (1).

3 Implementation details

All experiments ran on a single NVIDIA RTX 6000 Ada Generation GPU. We first trained vanilla FACT on SICS-155 and obtained about 82% frame accuracy. The boundary-aware model was then initialized from this checkpoint in a non-strict manner: shared weights were loaded, and the boundary heads were newly initialized. We used M=36 action tokens to stay compatible with the warm start. Inference for the boundary model is identical to the baseline FACT.

We tuned hyperparameters with W&B sweeps. For the baseline, the search covered learning rate $\{1\times10^{-4}, 3\times10^{-4}, 1\times10^{-3}\}$, batch size $\{4, 8\}$, temporal stride $\{1, 2, 3\}$, number of tokens $\{32, 36, 48\}$, matching strategy {o2o, o2m}, merge weight $\{0.25, 0.50, 0.75\}$, learning-rate decay epochs $\{40, 80, 120\}$, attention layers $\{4, 6\}$, and smoothing weight $\{0.0, 2.5, 5.0\}$.

For the boundary-aware variant, we focused on the boundary components: boundary loss weight $\{1.0, 1.5, 2.0\}$, TV exponent $\{2.0, 3.0\}$, and smoothing weight $\{1.0, 2.5\}$, while fixing the learning rate at 1×10^{-4} and the merge weight at 0.50. The best setting used boundary loss 1.0, TV exponent 2.0, smoothing 5.0, and M=36 tokens with the warm start described above.

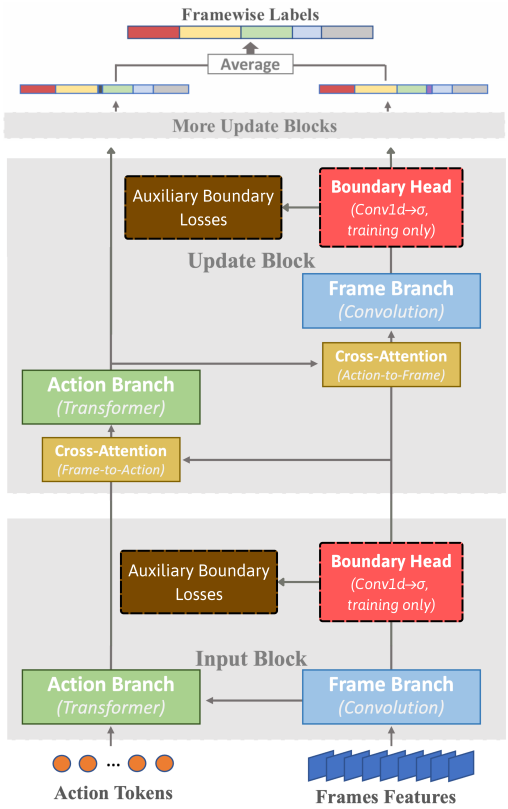

Fig. 1. Boundary-aware FACT. A boundary head (Conv1D→ σ) runs in parallel to the frame branch and is used only during training to supervise an auxiliary boundary BCE loss and to gate temporal smoothing (Eq. 3). Inference is unchanged and follows Eq. 1.

4 Results

Training uses the 100 videos in the SICS–155 training set, and evaluation is conducted on the 15 held-out test videos. The tuned FACT baseline attains 82.8% frame accuracy, 77.1 F1@0.50, and 86.9 Edit. Adding the boundary head improves accuracy to 84.1% and F1@0.50 to 78.6, with a small change in Edit to 86.3. In absolute terms, the boundary-aware variant yields +1.3 points Accuracy and +1.5 points F1, alongside slightly lower Edit (–0.6). Figure 2 illustrates the qualitative effect on a representative test video, showing cleaner transitions and fewer short spurious segments.

SICS–155 (19 phases) and SICS–105 (20 phases) [5] differ in size, number of phases. The external numbers therefore serve only as a reference band rather than a controlled baseline. Under those caveats, our boundary-aware FACT is

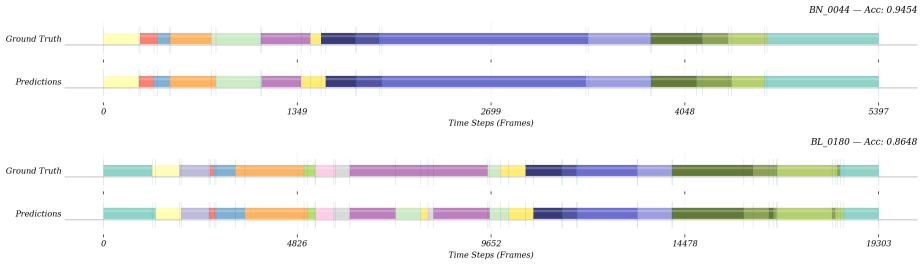

Fig. 2. Ground truth (top) versus predictions (bottom) on SICS–155. Boundaries align more cleanly with fewer short segments after adding the boundary-aware training.

Table 1. Validation on SICS–155 and an external point of reference from SICS–105 with 20 phases. The external row is not directly comparable due to different datasets and splits, but provides context.

Dataset & Method	#Phases	Acc (%)	F1@0.50	Edit
SICS–155 FACT (baseline)	19	82.8	77.1	**86.9**
SICS–155 FACT + boundary (proposed)	19	**84.1**	**78.6**	86.3
SICS–105 MS–TCN++ (Müller et al., Table 3)	20	80.87	75.02	78.27

competitive in Accuracy and F1 relative to the SICS–105 MS–TCN++ report, and substantially higher in Edit on our split.

5 Discussion and Conclusion

The proposed boundary-aware training improves a cross-attentional segmenter without changing the deployed model. A strategically placed boundary head learns a class-agnostic transition probability and gates the temporal-variation penalty during optimization. This keeps predictions stable within segments while allowing sharp changes at true transitions, reducing short spurious segments and yielding cleaner, better-calibrated trajectories. The extra cost is minimal because the head is a single Conv1D per block used only during training, and inference matches the baseline in parameters, memory, and runtime.

There are limits to this study. Gains depend on the training schedule and the learning-rate policy, and were most reliable with warm starts from a strong FACT checkpoint and a merge weight near 0.50. The supervision is binary and class-agnostic, which may under-represent ambiguous or gradual transitions. Results are reported on SICS-155 with I3D features; broader validation across backbones and datasets remains to be done.

For future work we will explore alternative ways to incorporate boundary awareness into the architecture, consider simple inference strategies that use boundary cues when helpful, and assess generality on additional datasets.

References

1. Lu, Z., Elhamifar, E.: FACT: Frame–Action Cross-Attention Temporal Modeling for Efficient Action Segmentation. In: CVPR (2024).
2. Ishikawa, Y., Kasai, S., Aoki, Y., Kataoka, H.: Alleviating Over-segmentation Errors by Detecting Action Boundaries. arXiv:2007.06866 (2020).
3. Farha, Y.A., Gall, J.: MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation. In: CVPR (2019).
4. Li, S., Farha, Y.A., Liu, Y., Cheng, M.-M., Gall, J.: MS-TCN++: Multi-Stage Temporal Convolutional Network for Action Segmentation. arXiv:2006.09220 (2020).
5. Mueller, S., Sachdeva, B., Prasad, S.N., *et al.*: Phase recognition in manual Small-Incision cataract surgery with MS-TCN++ on the novel SICS-105 dataset. Scientific Reports 15, 16886 (2025). https://doi.org/10.1038/s41598-025-00303-z
6. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal Convolutional Networks for Action Segmentation and Detection. In: CVPR (2017).
7. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical Transformer with Hierarchical Temporal Attention for Surgical Phase Recognition. arXiv:2408.03867 (2024).
8. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. arXiv:2303.16727 (2023).