Harnessing Temporal Information for Precise Frame-Level Predictions in Endoscopy Videos

Pooya Mobadersany[1(✉)], Chaitanya Parmar[1], Pablo F. Damasceno[1], Shreyas Fadnavis[1], Krishna Chaitanya[1], Shilong Li[1], Evan Schwab[2], Jaclyn Xiao[3], Lindsey Surace[1], Tommaso Mansi[1], Gabriela Oana Cula[1], Louis R. Ghanem[1], and Kristopher Standish[1(✉)]

[1] Janssen R&D, LLC, a Johnson & Johnson Company
{pmobader, kstandis}@its.jnj.com
[2] Epic Sciences, San Diego, CA, USA
[3] University of California, San Francisco, CA, USA

Supplementary

Segment	mAvg AUC (%)	mAvg F1 (%)	Accuracy (%)	Adj. accuracy (%)
IL	97.2 ± 0.0	84.7 ± 0.1	80.8 ± 0.1	93.1 ± 0.1
RC	92.1 ± 0.1	62.5 ± 0.2	63.8 ± 0.2	93.2 ± 0.1
TC	87.4 ± 0.1	53.2 ± 0.2	50.7 ± 0.2	94.8 ± 0.1
LC	91.0 ± 0.0	73.1 ± 0.1	75.3 ± 0.1	97.2 ± 0.1
RM	96.7 ± 0.0	79.0 ± 0.1	82.3 ± 0.1	98.2 ± 0.1

Table S1. Frame-level performance of EndoFormer for each anatomic segment on CD test set.

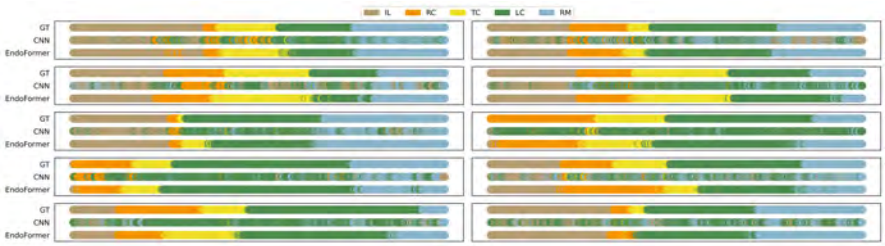

Fig. S1. Qualitative results on ten example endoscopy videos in the CD test set; each with top row: Ground Truth (GT) segments, middle row: CNN-based (CNN) predictions; bottom row: EndoFormer predictions. EndoFormer is doing a great job in predicting fully connected segments similar to GT. CNN, which is the Endomapper model adopted from [**?**] on the other hand fails to reconstruct fully connected segments, because of the lack of local and global temporal information.