Limited Scalability of Single Deep Neural Network for Surgical Instrument Segmentation in Different Surgical Environments

Daichi Kitaguchi, Hiro Hasegawa, Nobuyoshi Takeshita, Masaaki Ito (National Cancer Center Hospital East, Chiba, Japan)

Abstract

Background:

- Clarifying the scalability of a deep learning-based surgical instrument segmentation network in diverse surgical environments is important in recognizing the challenges of over-fitting in surgical device development.

Objectives:

- Comprehensively evaluate the scalability of a single deep neural network for surgical instrument segmentation.

Training set:

- The dataset for deep learning model training
  - 94 cases of lap-colorectal resection
  - 4,419 annotated images for T1: Harmonic, T2: Spatula-type electrode, T3: Forceps
  - Recorded with Olympus laparoscopic system

Test sets:

- Deep learning-based surgical instrument segmentation tasks were performed for the following test sets.
  - **Test set-1:** same condition as the training set
  - **Test set-2:** same recognition target surgical instrument and type of surgery but videos were recorded with different laparoscopic systems
  - **Test set-3:** same laparoscopic recording system and type of surgery, but a slightly different type of recognition target laparoscopic surgical forceps
  - **Test set-4:** same laparoscopic recording system and recognition target surgical instrument but different surgery type

*Abbreviation:
(m)AP: (mean) average precision, IoU: intersection over union
Ref.: reference value, Lap-: laparoscopic

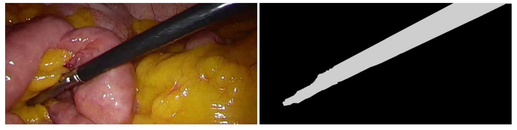
Test-1: Same Condition

Recognition Target Surgical Instruments

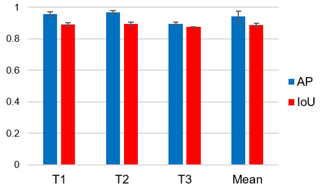


T1: Harmonic, T2: Spatula-type electrode, T3: Forceps

Annotation for Surgical Instruments



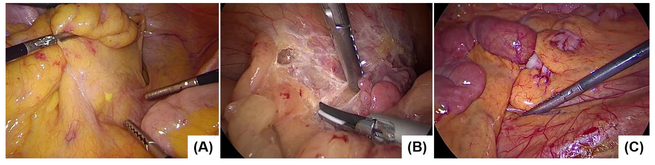
Raw image　　Annotated label

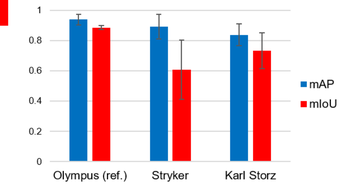


	AP	IoU
T1	0.958	0.892
T2	0.969	0.895
T3	0.895	0.876
Mean	0.941	0.887

Test-2: Different Types of Recording System

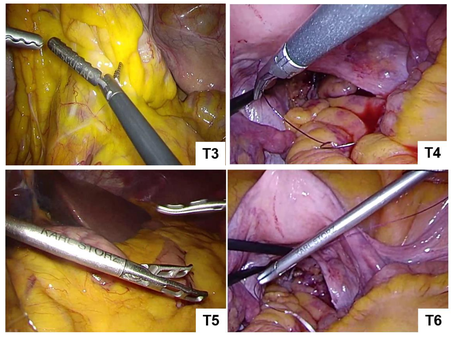
OLYMPUS　stryker　STORZ KARL STORZ—ENDOSKOPE



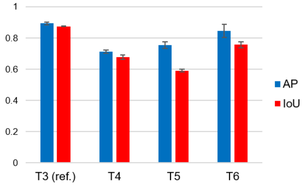


	mAP	mIoU
Olympus (ref.)	0.941	0.887
Stryker	0.893	0.608
Karl Storz	0.839	0.735

Test-3: Different Types of Forceps

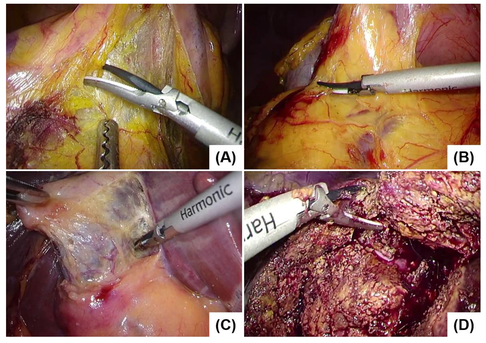


T4: Maryland, T5: Croce-olmi, T6: Needle holder

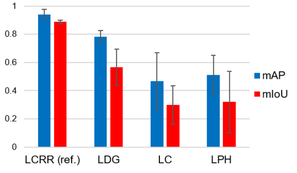


	AP	IoU
T3 (ref.)	0.895	0.876
T4	0.715	0.678
T5	0.756	0.592
T6	0.846	0.758

Test-4: Different Types of Surgery



(A): LCRR, (B): LDG, (C): LC, (D): LPH



	mAP	mIoU
LCRR (ref.)	0.941	0.887
LDG	0.782	0.565
LC	0.468	0.300
LPH	0.513	0.319

*Abbreviation:
LCRR: lap-colorectal resection, LDG: lap-distal gastrectomy, LC: lap-cholecystectomy, LPH: lap-partial hepatectomy

Conclusion

- The recognition accuracy significantly deteriorated even under slightly different conditions.
- To enhance the generalization of deep neural networks in the surgical field, it is crucial to construct a training set that considers the diversity of the surgical environment in a real-world setting.