Attention YOLACT++ for real-time instance segmentation of medical instruments in endoscopic procedures

Juan Carlos Angeles Ceron[1], Leonardo Chang[1], Gilberto Ochoa Ruiz[1], Sharib Ali[2]
[1]Tecnologico de Monterrey, School of Engineering and Sciences, Mexico
[2]University of Oxford, Institute of Biomedical Engineering, United Kingdom
A01271549@itesm.mx, gilberto.ochoa@tec.mx

Abstract

Image-based tracking of laparoscopic instruments plays a fundamental role in computer and robotic-assisted surgeries by aiding surgeons and increasing patient safety. Computer vision contests, such as the 2019 Robust Medical Instrument Segmentation (ROBUST-MIS) Challenge, encourage the development of robust models for surgical instrument segmentation and provide large, diverse, and extensive annotated datasets. To date, most of the existing models for instance segmentation of medical instruments were based on two-stage detectors, which provide robust results but are nowhere near to the real-time (5 frames-per-second (fps) at most). However, in order for the method to be clinically applicable, real-time capability is utmost required along with high accuracy. In this paper, we propose the addition of attention mechanisms to the YOLACT architecture that allows real-time instance segmentation of instrument with improved accuracy on the ROBUST-MIS dataset. Our proposed approach outperforms the winner of the 2019 ROBUST-MIS challenge in terms of robustness scores, obtaining 0.338 MI_DSC and 0.383 MI_NSD, while achieving real-time performance (37 fps).

1. Introduction

Computer-assisted minimally invasive surgery such as endoscopy has grown in popularity over the past years. However, due to the nature of these procedures, issues like limited view, extreme lighting conditions, lack of depth information and difficulty in manipulating operating instruments demand strenuous amounts of effort from the surgeons [6]. Surgical data science applications could provide physicians with context-aware assistance in order to overcome these limitations and increase the patient's safety. One of the main forms of assistance is by providing accurate tracking of medical instruments, as it is a fundamental prerequisite for tasks ranging from surgical navigation, skill analysis and complication prediction [3]. Nonetheless, accurate tracking of instruments often face difficult image scenarios such as bleeding, over/under exposure, smoke and reflections [4].

Computer vision contests, such as the Robust Medical Instrument Segmentation (ROBUST-MIS) Challenge [6] represent necessary efforts to encourage the development of robust models for surgical instrument segmentation. They integrate the developments in computer-assisted surgeries, and benchmark the generalization capabilities of the developed methods on different clinical scenarios. Furthermore, they provide large high-quality datasets to overcome one of the main bottlenecks of the development of robust methodologies, which is the lack of annotated data.

Previous approaches for instance segmentation of medical instruments were exclusively based in two-stage detectors such as Mask R-CNN [8]. While these models presented robust performance, they all suffer from high inference times, preventing them for achieving real-time performances, averaging around 5 fps. Realistically, real-time performance is mandatory in order to fully exploit the capabilities of tracking applications in live surgeries.

To overcome the speed bottleneck while maintaining a robust performance in terms of tool segmentation results, we propose a new approach based on YOLACT++ [1] equipped with attention modules on the multi-scale outputs of the network's backbone and Feature Pyramid Network (FPN). The increased representation power achieved by using attention allows the extraction of more discriminant features, suppressing the less effective ones and helping the model to learn salient features from the input images, which is essential for a robust performance. We carried out experiments using two different types of attention modules: Criss-cross Attention [7] and Convolutional Block Attention Modules [11]. The former, recursively integrates global context along feature maps in a fast and clever criss-cross fashion. The latter emphasizes relevant features along the channel and spatial axes by blending cross-channel and spatial information together. Our proposed model outperforms

others in the state of the art by a good margin, as it will be discussed in the rest of the paper.

2. Methods

2.1. Dataset

The Heidelberg Colorectal Data Set [3] served as a basis for the ROBUST-MIS challenge. It comprises 30 surgical procedures from three different types of surgery and includes detailed segmentation maps for the surgical instruments in more that 10,000 laparoscopic video frames. The generalizability and performance of the submitted algorithms was assessed in three stages with increasing levels of difficulty. In Stage 1, test data was taken from the procedures from which the training data were extracted. In Stage 2, test data was taken from the exact same type of surgery as the training data but from procedures not included in the training. Finally, in Stage 3, test data was taken from a different but similar type of surgery compared to the training data. A total of 996 frames with no visible instruments were removed from the training set, leaving 4,987 usable frames. From this subset, an 85-15 percent split was made for training and validation purposes respectively. As an additional step, the training and validation datasets were converted to COCO-style for easier integration with the YOLACT framework. We heavily applied data augmentation techniques to introduce as much variability as possible and increase the model's performance. The augmentation techniques that were used are random photometric distort, random scaling, random sample crop and random mirror.

2.2. Metrics

Two metrics were chosen to assess performance of the multiple instance segmentation task: Multiple Instance Dice Similarity Coefficient (MI_DSC) and Multiple Instance Normalized Surface Dice (MI_NSD). The DSC [2] is defined as the harmonic mean of precision and recall:

$$DSC(Y, \hat{Y}) := \frac{2 \mid Y \cap \hat{Y} \mid}{\mid Y \mid + \mid \hat{Y} \mid}, \qquad (1)$$

Where Y indicates the ground truth annotation and $\hat{Y}$ the corresponding prediction of an image frame.

The NSD measures the overlap of two mask borders [5]. The metric uses a threshold that is related to the inter-rater variability of the annotators. According to [6], their calculations resulted in a threshold of $\tau := 13$ for the challenge's data set.

To calculate the MI_DSC and MI_NSD, matches of instrument instances were computed. Then, the resulting metric scores per instrument instance per image were aggregated by the mean.

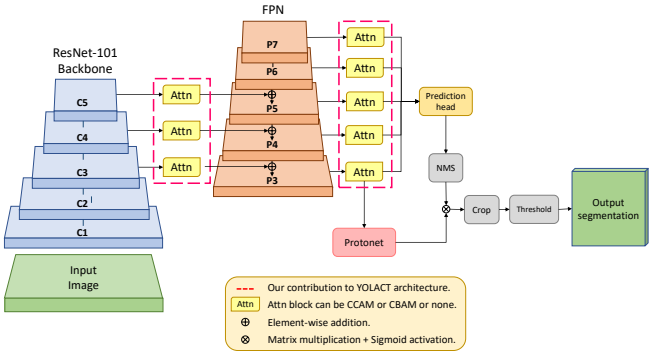

Figure 1. Proposed Attention YOLACT++ architecture with ResNet-101 backbone + FPN. Attention modules can be either CBAM or CCAM (which are removed in some experiments.)

Note that the challenge reports robustness and accuracy rankings. However, to compute accuracy it is mandatory to know the per image results per participant, which are not available due to privacy issues. For this reason, we will be reporting only robustness rankings.

The robustness rankings pay particular attention in stage 3 of the challenge and focus on the worst-case performance. For this reason, MI_DSC and MI_NSD are aggregated by the 5% percentile instead of by the mean or median [6].

2.3. Proposed model

In order to improve the robustness of the real-time YOLACT architecture used in our proposal, we introduce attention modules between the multi-scale outputs of the ResNet-101 backbone and the input of the FPN as well as on the output features of the FPN (see Figure 1). Attention allows the network to focus on the most relevant features without the need of additional supervision, enabling it to exploit salient features and avoid redundant use of information. We performed experiments with two types of attention modules: Criss-cross Attention (CCAM) [7] and Convolutional Block Attention (CBAM) [11]. A characteristic that these mechanisms have in common is that both are fast and computationally efficient, which is crucial in order to introduce as less time-processing overhead as possible.

CBAM sequentially infers a 1D channel and a 2D spatial attention maps which are aggregated to create refined features. The channel attention module extracts the inter-channel relationship of features by first aggregating spatial information through a combination of two pooling operations, generating two spatial context descriptors which are then forwarded to a multi-layer perceptron to create the channel attention map $M_c \in \mathbb{R}^{C \times 1 \times 1}$. Furthermore, the spatial attention module generates a spatial attention map by applying the same two pooling operations generating two 2D maps which are then concatenated and convolved to produce the 2D spatial attention map $M_s \in \mathbb{R}^{1 \times H \times W}$.

CCAM captures global contextual information efficiently. For each pixel in a feature map, it aggregates contextual information in its horizontal and vertical directions. By serially stacking two CCAM modules, each pixel can collect contextual information from all pixels in a given feature map. Next, the contextually rich feature is concatenated with the original feature maps, followed by one or several convolutional layers with batch normalization and activation for feature fusion. For each type of attention, we systematically attach modules first in the backbone's features, next in the FPN's features and finally on both locations (which we refer as *Full*), following the naming convention: *AttentionType-AttentionLocation*.

2.4. Experimental setup

Training was performed in an NVIDIA DGX-1 system. We trained the models for up to 100,000 iterations with a learning rate of 0.001, momentum of 0.9, weight decay of 5×10^{-4}, and batch size of 16. The performance was assessed using the evaluation code for the challenge [9] and the rankings were computed using the provided R package *challengeR* [10].

3. Results and discussion

Among our models, those based on CBAM achieved better performance than the ones based on CCAM. We hypothesize that CBAM can extract stronger representations as it generates attention maps for both the channel and spatial dimensions, unlike CCAM which only integrates spatial context. Regardless, attention-integrated models always outperformed the attentionless baseline in terms of robustness. Figure 2 shows dot-and-boxplots of the metric values for each algorithm over all test cases in stage 3 of the challenge. We can observe that YOLACT++ plus ResNet-101 is a strong baseline on its own, however, adding attention mechanisms boosts the performance particularly on instances below the second quartile, which are the most important for our performance metrics.

Next, we compare our top performing models and baseline to the top participants of the 2019 challenge (note that the 2020 edition did not take place). Table 1 shows the aggregated MI_DSC and MI_NSD values achieved for each participant/model. We achieve competitive results to the top performing methods, and in the case of *CBAM-Full* (CBAM on backbone and FPN) we outperform it by a significant margin on both metrics, while attaining a frame rate of 37 fps. Our results prompt to a high effectiveness of incorporating attention mechanisms to empower the learning capabilities of segmentation algorithms.

We observe high quality and temporally consistent masks. Figure 3 illustrates some examples with varying types and number of instruments. The model is robust to occluded instruments and various harsh conditions, like presence of blood, smoke, and poor lighting. Nevertheless, it struggles with transparent instruments and small instruments on the edge of the field of view. These problems will be addressed in future work.

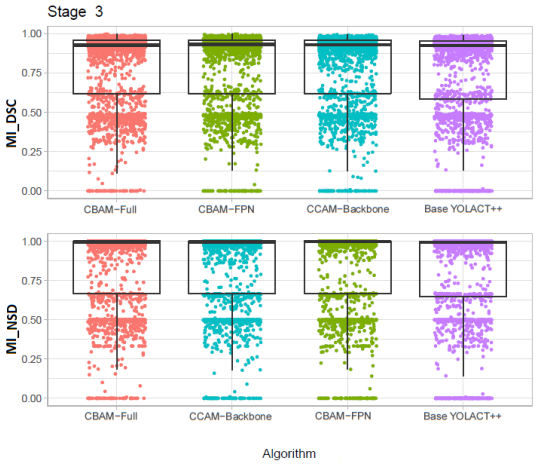

Figure 2. Dot-and-boxplots showing the performance of algorithms for every test case in stage 3 of the challenge. Plots were generated using the package challengeR [10].

Table 1. Aggregated evaluation performance for stage 3 of the challenge by 5% percentile. The top section of the table shows the 3 best performing teams (2019 Edition); scores were taken from [6]. The bottom section includes our best 3 experiments plus the baseline model, which does not make use of attention modules.

Team/Algorithm	MI_DSC	MI_NSD	FPS
www	0.31	0.35	5*
Uniandes	0.26	0.29	5*
SQUASH	0.22	0.26	5*
CBAM-Full	**0.338**	**0.383**	37
CBAM-FPN	0.315	0.333	37
CCAM-Backbone	0.313	0.338	31
Base YOLACT++	0.000	0.000	**43**

*Approximated from their Mask R-CNN base, real measurement is not reported

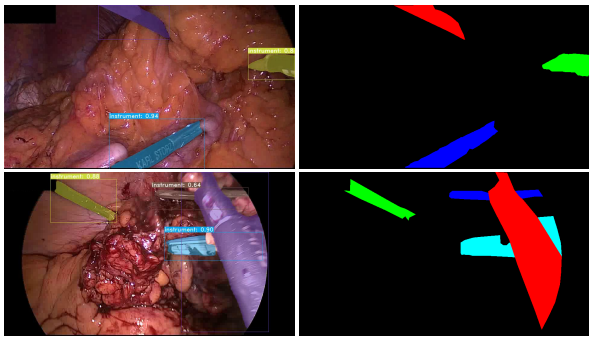

Figure 3. Side-by-side comparison of *CBAM-Full* evaluation results and ground truth annotations.

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020. 1

[2] Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 2

[3] Lena Maier-Hein *et al.* Heidelberg colorectal data set for surgical data science in the sensor operating room, 2021. 1, 2

[4] Sebastian Bodenstedt *et al.* Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery, 2018. 1

[5] Stanislav Nikolov *et al.* Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy, 2021. 2

[6] Tobias Ross *et al.* Robust medical instrument segmentation challenge 2019, 2020. 1, 2, 3

[7] Zilong Huang *et al.* Ccnet: Criss-cross attention for semantic segmentation, 2020. 1, 2

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018. 1

[9] Tobias Roß and Annika Reinke. Robustmis2019. 3

[10] Manuel Wiesenfarth, Annika Reinke, Bennett Landman, Manuel Jorge Cardoso, Lena Maier-Hein, and Annette Kopp-Schneider. challenger: Methods and open-source toolkit for analyzing and visualizing challenge results. 3

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module, 2018. 1, 2