Medical Image Analysis 81 (2022) 102569

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Real-time instance segmentation of surgical instruments using attention and multi-scale feature fusion

Juan Carlos Ángeles Cerón [a], Gilberto Ochoa Ruiz [a,*], Leonardo Chang [a], Sharib Ali [b,c,d,**]

[a] Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, Mexico
[b] Institute of Biomedical Engineering (IBME), Department of Engineering Science, University of Oxford, Oxford, UK
[c] Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK
[d] School of Computing, University of Leeds, Leeds, UK

ARTICLE INFO

Keywords:
Deep learning
MIS instance segmentation
Real-time
Single-stage
Attention
Multi-scale feature fusion

ABSTRACT

Precise instrument segmentation aids surgeons to navigate the body more easily and increases patient safety. While accurate tracking of surgical instruments in real-time plays a crucial role in minimally invasive computer-assisted surgeries, it is a challenging task to achieve, mainly due to: (1) a complex surgical environment, and (2) model design trade-off in terms of both optimal accuracy and speed. Deep learning gives us the opportunity to learn complex environment from large surgery scene environments and placements of these instruments in real world scenarios. The Robust Medical Instrument Segmentation 2019 challenge (ROBUST-MIS) provides more than 10,000 frames with surgical tools in different clinical settings. In this paper, we propose a light-weight single stage instance segmentation model complemented with a convolutional block attention module for achieving both faster and accurate inference. We further improve accuracy through data augmentation and optimal anchor localization strategies. To our knowledge, this is the first work that explicitly focuses on both real-time performance and improved accuracy. Our approach out-performed top team performances in the most recent edition of ROBUST-MIS challenge with over 44% improvement on area-based multi-instance dice metric MI_DSC and 39% on distance-based multi-instance normalized surface dice MI_NSD. We also demonstrate real-time performance (> 60 frames-per-second) with different but competitive variants of our final approach.

1. Introduction

Surgical site infection (SSI) has been the most common cause of hospital-acquired infection and the most common way of infection transmission in patients undergoing surgery (Caroff et al., 2019). It is therefore imminently important to develop strategies for reducing such infection rates. Minimally invasive surgical (MIS) procedures compared to open surgery lowers such risks. For these reasons and due to the growth the data science in the operating room applications, there exists an increasing demand for computer assisted surgery to improve the efficacy of MIS (Bartoli et al., 2012; Sheetz et al., 2020).

Computer-assisted minimally invasive surgery methods such as endoscopy have grown in popularity in recent years. However, due to the nature of these procedures, issues like limited field-of-view, extreme lighting conditions, lack of depth information and difficulty in manipulating operating instruments demand strenuous amounts of effort from the surgeons (Roßet al., 2021). Surgical data science applications (Maier-Hein et al., 2021a) could provide physicians with context-aware assistance during minimally invasive surgery in order to overcome these limitations and increase patient safety.

One of the main forms of assistance is by providing accurate tracking of medical instruments using computer vision (CV) techniques, such as object detection and localization or instance segmentation methods (Ward et al., 2021). These systems are expected to be a crucial component in tasks ranging from surgical navigation, skill analysis and complication prediction during surgeries, as well as other computer integrated surgery (CIS) applications (Maier-Hein et al., 2021b; Fu et al., 2021). Nonetheless, methods for accurate tracking of instruments are often deployed in difficult operational scenarios in which the presence of bleeding, over or under exposed frames, smoke, reflection and other types of artifacts are oftentimes unavoidable (Bodenstedt et al., 2018). The net effect of these issues increases the missed detection rates in endoscopic surveillance, limiting the overall robustness of CV algorithms, hampering the adoption of AI-based tools in this context (Ali et al., 2021). Moreover, real-time deployment of such

* Corresponding author.
** Corresponding author at: School of Computing, University of Leeds, Leeds, UK.
E-mail addresses: gilberto.ochoa@tec.mx (G.O. Ruiz), s.s.ali@leeds.ac.uk (S. Ali).

https://doi.org/10.1016/j.media.2022.102569
Received 9 November 2021; Received in revised form 1 July 2022; Accepted 4 August 2022
Available online 6 August 2022

tools is of tremendous value and one of the major requirement for it to be applied in clinical settings. Therefore, the development of robust and real-time techniques that can be effectively deployed during real endoscopic interventions is of utmost importance. In this regard, recent years have seen a significant increase in the number of CV contests geared towards endoscopy. More specifically, the Robust Medical Instrument Segmentation (ROBUST-MIS) Challenge (Roßet al., 2021) at the International Conference on Medical Image Computing and Computer Assisted Interventions (MICCAI) has sought to address some of the issues discussed above. This challenge represents an important and necessary effort to encourage the development of robust models for surgical instrument segmentation, integrating the developments in computer-assisted surgeries, and as a benchmark for the generalization capabilities of the developed methods on different clinical scenarios. Furthermore, the challenge organizers provide a large high-quality dataset in an effort to overcome one of the major bottlenecks of the development of robust methodologies, i.e., the lack of annotated multi-instance instrument segmentation data. The development of surgical tool navigation and tracking methods in a complex environment will enable improved patient care during surgery by maximizing the focus of surgeons, accelerating research in CIS as well.

Previous approaches for instance segmentation submitted to the ROBUST-MIS challenge have been mostly based on two-stage detectors such as Mask R-CNN (He et al., 2018). While these models exhibited decent performances in terms of robustness, they suffer from high inference times due to well-known architectural limitations of such models, preventing them from achieving real-time performance (i.e. mean average test time of only around 5 frames-per-second, fps). However, real-time performance is mandatory in order to fully exploit the capabilities of tracking applications in surgeries. While deep learning methods using lightweight models are also available, they fail to robustly segment objects in endoscopy imaging (Ali et al., 2020). Similarly, there is a trend of using ensemble models to improve the overall segmentation accuracy in these images. However, combining few models together drastically increases the inference time, thereby making it less feasible to deploy such models in clinical settings (Xu et al., 2020).

In order to overcome the current inference limitations, while maintaining a robust performance in terms of tool segmentation results, we propose a new approach based on the single-stage model for instance segmentation. Although recent years have seen a steady increase in the search for more capable one stage detectors and instance segmentation architectures (for example PolyYolo Hurtik et al., 2020, BlendMask Chen et al., 2020 and Solov2 Wang et al., 2020b), the majority of these models have not been used in the context of endoscopic computer vision due to the performance gap they still present. The YOLACT++ (Bolya et al., 2020) architecture is one of the most recent methods for real-time instance segmentation and it is particularly appealing due to its simplified architecture capable of learning to localize instance masks automatically with minimum computational overhead. It does so by generating a dictionary of non-local prototype masks over the entire image and predicting a set of linear combination coefficients per instance. Thus, for this contribution we have used YOLACT++ as a baseline architecture upon which we have developed several improvements to make it suitable for robust surgical instrument instance segmentation. To this extend, we have explored the use of attention modules on the outputs of the network's backbone and feature pyramid network (FPN) at multiple scales. Moreover, we have additionally carried out a series of optimization techniques by analyzing the worst-performing frames of the best model in our experiments. The optimization techniques includes domain-specific data augmentation, anchor optimization, and multi-scale feature fusion. The main contributions of our work are summarized as follows:

- A real-time single-stage instance segmentation framework with attention mechanism.
- Exploration of domain-targeted data augmentation techniques tailored to the ROBUST-MIS challenge dataset
- Anchor box optimization via a differential evolution search (Zlocha et al., 2019) for the ROBUST-MIS Challenge.
- Integration of global contextual features by deploying a multi-scale fusion block in the network's backbone
- A thorough analysis of the worst-case samples to evaluate each of the tested model

The rest of the paper is organized as follows. In Section 2 we present previously published work related to medical instrument segmentation, instance segmentation methods, attention mechanisms and multi-scale feature fusion network. In Section 3, we present the details of the ROBUST-MIS dataset and our proposed approach for surgical instrument segmentation. Section 4 presents our data preparation, experimental setup and results on ROBUST-MIS dataset. For completeness, we also provide results of our proposed method on the EndoVis 2017 instrument segmentation dataset (Allan et al., 2019). In Section 5, we discuss the effects of the different types of network configurations and propose future directions. Finally, Section 6 concludes the paper.

2. Related work

In this section, we will discuss some of the most important aspects to understand the proposed contribution, namely: instance segmentation and its current limitations, recent works in attention mechanisms, anchor box optimization techniques specifically tailored to the addressed problem and finally, multi-scale fusion networks that followed to make our extended instance segmentation model more robust.

2.1. Deep learning for instrument segmentation

Deep learning has accelerated research for surgical instrument segmentation and the public access of labeled data via instrument segmentation challenges mostly at EndoVis (refer to Allan et al., 2019, 2020, Roßet al., 2021) have contributed to these developments over recent years. Built upon the UNet model (Ronneberger et al., 2015a), LinkNet and TernausNet were developed for instrument segmentation on robotic surgery datasets (Shvets et al., 2018), acquired by da Vinci Xi surgical system of several different porcine procedures made available in EndoVis17 (Allan et al., 2019). Milletari et al. (2018) proposed a convolutional long short term memory (LSTM) with deep residual networks using a coarse-to-fine strategy showing greater improvements over other state-of-the-art approaches including UNet (Ronneberger et al., 2015a) and FCN (Long et al., 2015a) on the EndoVis 2015 instrument segmentation challenge dataset[1] focused on laparoscopic and robotic surgery.

The recent "Robotic Instrument Segmentation Sub-Challenge" introduced also at EndoVis was oriented towards binary segmentation, subcomponent segmentation and instrument identification and instance segmentation tasks (Roßet al., 2021). The challenge was geared towards assessing the robustness and generalization capabilities of the deep learning models. Most of the competing methods in this challenge have been mostly based on Mask-RCNN (He et al., 2017) implementations and its variants for the multi-class instance segmentation and binary segmentation tasks. Participants also explored methods such as OR-UNet (Isensee and Maier-Hein, 2020), DeepLabV3+ (Chen et al., 2018), U-Net (Ronneberger et al., 2015b) and RASNet (Ni et al., 2019). The best performing methods for the binary segmentation task were OR-UNet and DeepLabv3+ with pre-trained ImageNet encoders. Some other contestants also explored the use of ensemble methods, but they were typically limited in speed, and thus some of the most

[1] https://endovissub-instrument.grand-challenge.org/EndoVisSub-Instrument/

robust methods are incapable of attaining the real-time performances required for realistically segmenting and tracking objects on endoscopic video data. A double decoder–encoder network was explored for faster binary mask segmentation on ROBUST-MIS'19 dataset (Jha et al., 2021) that outperformed several state-of-the-art methods. Recently, a one-shot instrument segmentation method (Zhao et al., 2021) using anchor guided meta-learning approach was proposed and validated on several publicly available datasets (including EndoVis18 (Allan et al., 2020) porcine data).

2.2. Real-time instance segmentation methods

While extensive research has been conducted for the development of real-time object detection and semantic segmentation models, few works have tackled the problem of real-time instance segmentation (He et al., 2017; Bolya et al., 2019). This is due to the increased complexity in the instance segmentation task that requires predictions of instance labels and pixel-level segmentation simultaneously. One-stage methods (Hurtik et al., 2020; Lee and Park, 2019; Chen et al., 2020) though conceptually faster than two-stage methods (e.g., Mask RCNN He et al., 2017), still require many non-trivial computations (e.g., mask voting). This severely limits their speed making them not suitable for real-time applications. In contrast, recent methods (Bolya et al., 2019, 2020) make use of lightweight assemblies of masks (only linear combinations are used), making the approach very efficient. Although YOLACT (Bolya et al., 2019) was one of the first real-time one-stage instance segmentation approaches, the accuracy gap compared to Mask R-CNN (He et al., 2017) was still significant. While Mask R-CNN is based on a two-stage object detector (e.g., Faster R-CNN Ren et al., 2015), YOLACT (Bolya et al., 2019) is built on one-stage detector (RetinaNet Lin et al., 2017b) that directly predicts boxes without a proposal step, limiting its accuracy. This was partially addressed with the introduction of Yolact++ (Bolya et al., 2020), which incorporated deformable convolutions into the backbone network, improving the feature sampling and yielding an improved accuracy. Furthermore, the prediction head was optimized with better anchor scale and aspect ratio choices for an increased object recall.

2.3. Attention mechanisms

Attention has been able to boost model performance across a wide range of computer vision tasks, such as image captioning (You et al., 2016), visual question answering (Xu et al., 2016), and visual attribute prediction (Seo et al., 2018). Attention allows the network to focus on the most relevant features without the need of additional supervision, preventing redundant use of information and extracting salient features that are useful for a given task. Attention mechanisms enable convolutional neural networks to overcome the size limitations of its receptive field, as it has proven to be excellent at extracting global dependencies between inputs and outputs, thus, improving the modeling of long-range dependencies even at opposite ends of and image (Sinha and Dolz, 2021). For example, similar textures may appear in different parts of an image, several disjoint semantic cues may provide insight to the general classification of an image, and an object might present complex and occluded parts throughout an image (Chaudhari et al., 2019). In the context of medical instance segmentation, multiple attention-based models (Kaul et al., 2019; Gu et al., 2020; Sinha and Dolz, 2021) have obtained state-of-the-art performance in fields like brain tumor, skin cancer and lung lesion segmentation on CT scans and X-rays, respectively. However, until now, instance segmentation of medical instruments in laparoscopic surgeries using attention mechanisms has not been fully explored.

Table 1
Training and test sample distribution for each both training and test stages of the ROBUST-MIS challenge (Roßet al., 2021). The quantities in parenthesis represent the % of frames with no instrument instance.

Procedure	Training	Testing		
		Stage 1	Stage 2	Stage 3
Procto-colectomy	2,943 (2%)	325 (11%)	255 (11%)	0
Rectal resection	3,040 (20%)	338 (20%)	289 (15%)	0
Sigmoid resection[a]	0	0	0	2,880 (23%)
TOTAL	5,983 (17%)	663 (15%)	514 (13%)	2,880 (23%)

[a]Unknown surgery.

2.4. Multi-scale feature fusion

Due to the wide range of scale variation of objects found in instance segmentation, multi-scale features are essential for robust performance (Wang et al., 2020a). Therefore, multi-scale feature aggregation is an adequate strategy to create detailed parsing maps (Ding et al., 2018). Current methods address this issue by using encoder–decoder architectures (Ronneberger et al., 2015a; Lin et al., 2017a) that combine high level and low level features at a single scale (Long et al., 2015b) or multiple scales (Ronneberger et al., 2015a). However, these approaches suffer from redundant use of information (Sinha and Dolz, 2021). In the field of medical instrument segmentation, the use of multi-scale aggregation has not been fully explored, especially combined with attention for improved robustness.

3. Materials and method

In this section we present details on the dataset used in our study and we describe our proposed framework for multi-instance surgical instrument segmentation in detail.

3.1. The ROBUST-MIS challenge dataset

For our experiments we made use of the Robust Medical Instrument Segmentation (ROBUST-MIS) (Maier-Hein et al., 2021b) challenge dataset which is the first large-scale annotated MIS dataset. The dataset is comprised of a total of 10,040 annotated video frames extracted from 30 minimally invasive daily-routine surgical procedures and includes detailed segmentation ground truth masks for the surgical instruments present in these frames. The surgical procedures include 10 rectal resection procedures, 10 proctocolectomy procedures, and 10 sigmoid resection procedures. The image resolution of all the provided frames is 960×540 pixels. In order to measure the robustness test, the dataset is comprised of three unique test sets and divided into different stages:

> **Stage 1:** Test data taken from the same procedures from which the training data were extracted
> **Stage 2:** Test data taken from the exact same type of surgery as the training data but from procedures (patients) not included in the training
> **Stage 3:** Test data taken from a different (unseen) but similar type of surgery and different (unseen) patients

The detailed training and test distribution is summarized in Table 1. However, for testing we have used only test set 3, *i.e. Stage 3*, which is from an unseen Sigmoid resection procedure and allows us to validate on the generalizability of the proposed framework directly. Sample images for challenging frames provided in the test set 3 are shown in Fig. 1.

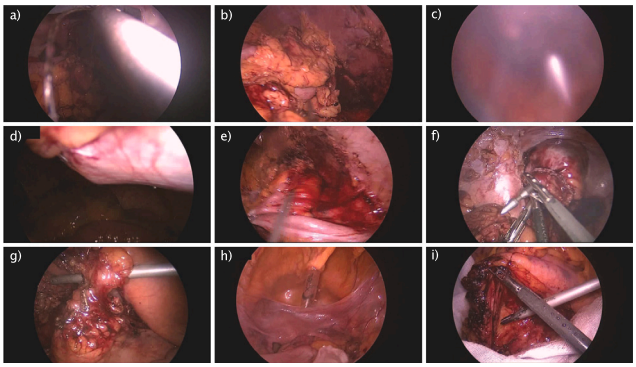


Fig. 1. Challenging images present in test dataset. Stage 3 test data consisted of several frames that included a) instrument flare, b) partial occlusions due to blood, c) occlusion due to smoke, d) underexposed regions with instrument, e) motion blur, f) multiple different instruments in the scene, g) partial occlusion due to organ, h) transparent instrument and i) different instruments crossing.

3.2. Method

This section describes our proposed approach for multi-instance segmentation of medical instruments in the ROBUST-MIS dataset. Our proposed framework (see Fig. 2) is built upon the single-stage YOLACT++ (Bolya et al., 2020) instance segmentation architecture. For our framework we employ ResNet-101 (He et al., 2015) as the backbone network, followed by a multi-scale feature fusion (MSFF) module used to aggregate contextual information from the feature maps across all scales (i.e. high-to-low resolution feature representations). Each of these contextually rich fused features are then passed through attention modules to further refine these representations before being forwarded to the feature pyramid network (FPN) (Lin et al., 2017c). A second set of attention modules are then applied to further enhance the FPN output features allowing for an improved performance of the prototype network and our anchor-optimized prediction head. Finally, we perform classical non-maximum suppression for final mask instance prediction which is then combined with the prototype mask and the cropping provides the predicted bounding box.

Below we describe our MSFF module, attention mechanisms, and the anchor optimization used our framework.

3.2.1. Multi-scale feature fusion

To aggregate multi-scale features, while maintaining a high-resolution representation, we integrated a fusion module inspired by the method proposed in Wang et al. (2018). Considering the features at different scales indicated as F_s where s denote the scale level in the architecture (see Fig. 3).

Features from each level s are up-sampled through transposed convolution to the size of the highest resolution feature maps in the architecture, leading to enlarged feature maps F'_s. Next, all F'_s are concatenated into a single tensor which is passed through a convolutional layer to integrate context from all scales into a single feature map $F_{MS} = conv([F'_0, F'_1, F'_2, F'_3, F'_4])$. In this manner, F_{MS} encodes both low-level and high-level semantics learned at different stages in the network. Finally, F_{MS} is concatenated with each of the F'_s feature maps and convolved to aggregate multi-scale information, creating multi-scale fused feature maps F_A.

Note that multi-scale feature fusion can be applied in the same way in both the backbone features and the FPN features. We opted to attach it on the backbone features as we believe it would create stronger representation which will get even further refined on the FPN.

3.2.2. Attention mechanisms

We employed Criss-cross Attention Modules (CCAM) (Huang et al., 2020) and Convolutional Block Attention Modules (CBAM) (Woo et al., 2018), specifically due to their fast and computationally efficient performance, which is paramount to introduce as less computational overhead as possible into the model, thus maintaining low inference times. We attach the attention modules between the backbone and neck structures, as well as the neck and head of our network (see Fig. 2).

The rationale behind the selection of these locations is that the addition of attention allows the model to extract richer context by aggregating local information with its corresponding global dependencies (Huang et al., 2020). Additionally, attention aids to emphasize interdependent relationships between channel maps and between spatial regions without additional supervision (Woo et al., 2018). Since the backbone network and the FPN are where most of the semantic context is distilled, it is a natural choice to attempt to refine their feature representations using attention, especially since mask prototypes and the prediction head benefit from better features.

3.2.3. Prototype generation

ProtoNet is the prototype generation branch in the architecture that predicts a set of k "prototype masks" for the entire image. It is implemented as a Fully Connected Network (FCN) in which the last layer predicts k-channels consisting of a set of image-sized masks. These masks do not depend on any one instance and the ProtoNet is itself attached to the backbone feature layer through a Feature Pyramid Layer (FPN) (see Fig. 2).

ProtoNet enables the generation of robust masks (from deeper features) with higher resolution resulting in a higher quality masks and improved performance for small object segmentation. Here, the largest feature layer P3 of a FPN network is first used to extract prototype masks which are then up-sampled to one fourth of the input image to increase performance on small objects. To produce instance masks, the generated maps of the prototype branch are combined with that of the instance maps from the prediction head, using a linear combination of the former with the latter as coefficients. Finally, a *sigmoid* non-linearity is applied to produce the final output instance masks.

3.2.4. Loss function and anchor optimization

We used a combination of three losses (Bolya et al., 2020), namely, classification loss L_{cls}, bounding box regression loss L_{box}, and mask loss L_{mask} for model training with weights of 1, 1.5 and 6.125, respectively. The individual weights are chosen such that they equally contribute to the final loss function L_{seg}. For L_{mask} the value of 6.125 is computed based on the number of prototype masks generated; this value is used to normalize the ProtoNet branch outputs to diminish the overpowering activation from each prototype (Bolya et al., 2020). For computing the mask loss we simply calculated the pixel-wise binary cross entropy between estimated masks $\mathbf{M}$ and the ground truth masks $\mathbf{M}_{gt}$: $L_{mask} = BCE(\mathbf{M}, \mathbf{M}_{gt})$. We employ *softmax* cross entropy for L_{cls} with one positive label (*i.e.*, instrument class) and a background label. For the L_{box} we used smooth-L_1 loss. An equally weighted (weight of 1) fourth loss referred to as semantic segmentation loss L_{sem} is used to improve feature richness on some layers that are evaluated only during training, where the ground truths for this loss is computed from instance annotations (Bolya et al., 2020). The provided weights for each component in L_{seg} were empirically set on baseline YOLACT++ method.

We experimented using different weight combinations in the loss function L_{seg} in the baseline model. Empirically we found that changing the weights from default values impacted negatively on the model performance. For example, setting the weights of all loss functions to a value of 1, resulted in a performance degradation of about 4.18%.

For improving the results in the ROBUST-MIS challenge, we further optimized the anchor boxes by deploying the same strategy as

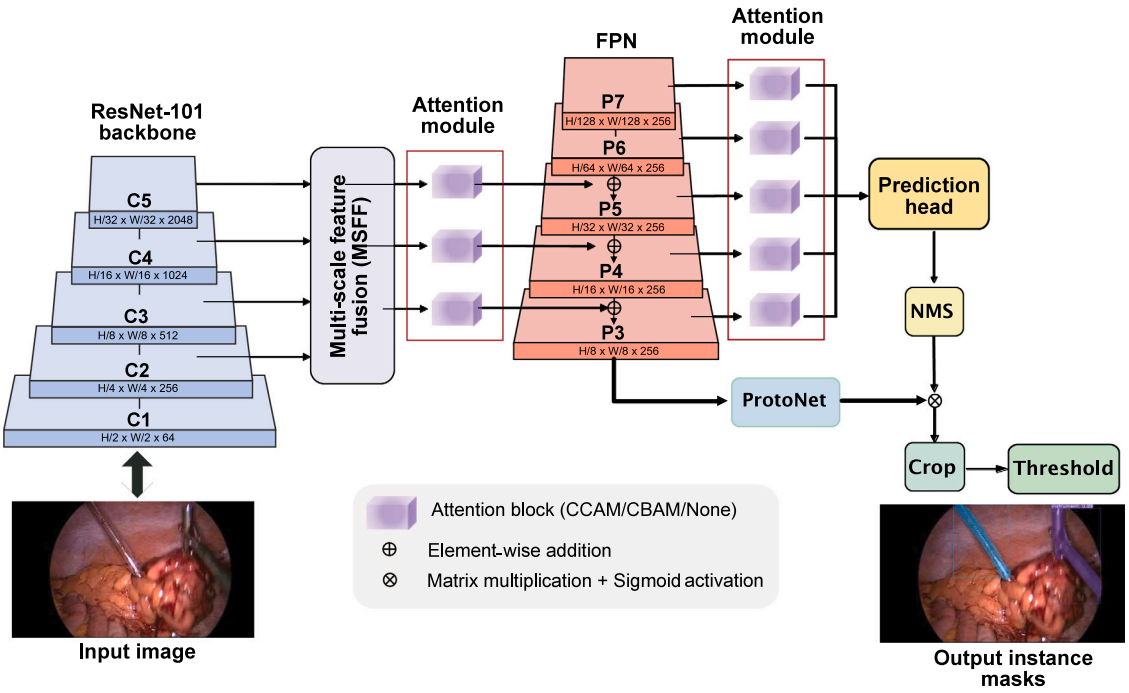


Fig. 2. Overview of our proposed framework. Our architecture is built upon popular single-stage YOLACT++ and comprises of additional multi-scale feature fusion and attention modules. It is to be noted that the attention modules can be easily interchangeable from criss-cross attention (CCAM) to convolutional block attention module (CBAM) or none.

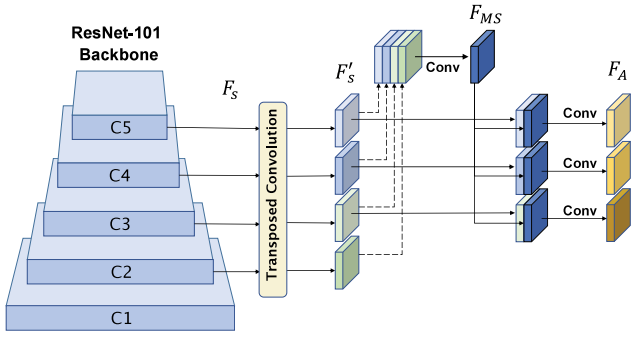


Fig. 3. Multi-scale feature fusion (MSFF) module. Feature maps from the different scale feature maps of the backbone are combined to create aggregated features from all scales.

in (Zlocha et al., 2019), in which a differential evolution search algorithm was used to optimize the scale and ratio of anchors in the validation set. We used the algorithm to find best anchor settings for 5 scales and 5 ratios (Ren et al., 2015; Zlocha et al., 2019) which was done by maximizing the overlap between the target instrument bounding-box and the best anchor on the validation dataset. The resulting values after running the anchor box optimization algorithm in the ROBUST-MIS dataset, we obtained [0.435, 0.502, 0.578, 0.664, 0.762] for scales and [0.267, 0.554, 1.0, 1.804, 3.746] for aspect ratios. We observed that three of the optimized ratios ([0.554, 1, 1.804]) were similar to the default values ([0.5, 1, 2]). Therefore, we opted to keep the default ratios that are rounded to one decimal place. Similarly, the remaining two ratios [0.267, 3.746] corresponding to the long horizontal and vertical objects, respectively, were approximated to [0.25, 4] leading to a final rounded aspect ratio of ([0.25, 0.5, 1.0, 2.0, 4.0]). It is to be noted that no test data were used in finding the optimal scale and ratio values reported in this work.

4. Experiments and results

4.1. Data preparation

As shown in Table 1, the ROBUST-MIS dataset contains about 17% empty frames (ef) on its training set. These frames do not have any visible instruments in them, and although we could have left them as negative examples for training, we opted to remove them from the training set. This decision was taken considering that the data already has plenty of negative examples in the frames' background for the model to learn from; additionally, removing such frames speeds up the training. In the end, a total of 996 images with no visible instruments were discarded, leaving 4,987 frames in our training set. We then applied an 85%–15% split to the curated training set with 15% for validation purposes. The training set was then randomly shuffled before creating the train and validation splits. We finally obtained a training set composed of 4,239 frames and a validation set comprised of 748 frames. As a final step, the training and validation datasets were converted to COCO-style format, which involved extracting mask contours and generating bounding box coordinates from the provided annotation images and translating them to the target JSON format.

4.2. Training setup

The training was performed on an NVIDIA DGX-1 system consisting of 8 NVIDIA Volta-based GPUs; however, each model was trained on *a single* GPU. The models were trained for up to 400,000 iterations with a learning rate of 0.001, momentum of 0.9, weight decay of $5e{-}4$, and a batch size of 16.

We applied data augmentation techniques to increase our model performance. These included random photometric distortions (i.e., changes in contrast, color-space, saturation, hue, brightness, and noise transformations) and affine transformations (i.e., random scaling and random sample crop). These augmentations were incrementally applied to the model to assess their effect on the model performance. As we noticed that some miss-detections were due to the presence of various rotations of the instruments, we further applied augmentations consisting of additional random flips and random rotations to the previous list.

Table 2
Model configurations for our experiments with integration of different attention mechanisms in unique settings.

Model identifier	Attention type	Backbone attention	FPN attention
Base YOLACT++			
CCAM-Backbone	CCAM	✓	
CCAM-FPN	CCAM		✓
CCAM-Full	CCAM	✓	✓
CBAM-Backbone	CBAM	✓	✓
CBAM-FPN	CBAM		✓
CBAM-Full	CBAM	✓	✓
CBAM-Full + Aug	CBAM	✓	✓
CBAM-Full + Aug + Anch	CBAM	✓	✓
CBAM-Full+ Aug + Anch MS	CBAM	✓	✓

4.3. Ablation study setup

Our experiments systematically integrate attention mechanisms in two strategic locations of the baseline YOLACT framework: (1) at the output of each convolutional block of the ResNet-101 backbone (He et al., 2015), and (2) at the multi-scale output features of the FPN (Lin et al., 2017c).

The incorporation of attention in these locations was alternated throughout experiments leading to three different network configurations:

1. Exclusively incorporated in the backbone
2. Exclusively incorporated in the FPN
3. Integrated in the backbone and FPN simultaneously (which we refer to as a *Full* configuration)

At the end, six attention-based models were created by following this strategy, plus a baseline network without attention. Next, we selected the top performing configuration from the six mentioned network configurations. We then applied other optimization techniques to understand the network performance that included domain-targeted data augmentation and anchor optimization. Finally, we added our multi-scale MSFF module. Table 2 summarizes all the different model configurations.

4.4. Metrics and assessment

The algorithms' performance was evaluated following the guidelines defined by the ROBUST-MIS Challenge. Robustness performance was assessed considering the area-based metric multi-instance dice MI_DSC and the distance-based multi-instance normalized surface dice MI_NSD using the code implementations employed in the challenge which are provided in Roß and Reinke (2019). Furthermore, our reported model rankings were computed using the publicly available *challengeR* (Wiesenfarth et al., 2021) R package developed by the challenge organizers to accurately evaluate competitors. The ranking stability was investigated using bootstrapping for quantifying ranking variability using 1000 samples.

The robustness rankings are particularly focused on models' capabilities in stage 3 of the challenge and pay particular attention on the worst-case performance of methods. For this reason, the robustness rankings are computed by aggregating the resulting scores for all the test cases by the 5% percentile instead of by the mean or median (Roßet al., 2021).

We performed inference speed assessments by running inference tests on a 10 s video snippet from the ROBUST-MIS dataset a total of ten times per model. The reported frame rates were then aggregated by the mean. Inference was tested on a *single* Tesla P100 GPU from the DGX-1 cluster with video multi-frame enabled.

4.5. Results

In this section, we compare our results to the ROBUST-MIS challenge methods and present both quantitative and qualitative results.

4.5.1. Quantitative results

Table 3 shows the detailed result for the multi-instance segmentation task on the ROBUST-MIS dataset. It can be observed that overall, attention-based models show improvement over the previous approaches for the aggregated MI_DSC and MI_NSD, and most notably FPS. *CCAM-Full*, *CCAM-Backbone*, and *CBAM-FPN* achieved competitive results in terms of MI_DSC compared to the top contestant *www*, scoring 0.30, 0.31, and 0.31 respectively against *www*'s 0.31. On the other hand, the three models fall behind by 0.02 in average regarding MI_NDS. Nonetheless, such a small difference in performance is outweighted by the dramatic increase of inference speed of at least 9× from all the models.

The highest metric scores from the initial ablation experiments correspond to *CBAM-Full*, which resulted in 0.34 MI_DSC and 0.38 MI_NSD. *CBAM-Full* presents an improvement of 2.8% on MI_DSC and 3.3% on MI_NSD compared to the previously best model while attaining real-time inference speed of 65 FPS. The additional domain-targeted data augmentation efforts applied in *CBAM-Full + Aug* resulted in an improvement of 4.4% and 4.6% on MI_DSC and on MI_NSD, respectively, with respect to *CBAM-Full*. *CBAM-Full + Aug + Anch* greatly benefited from anchor optimization, which resulted in the model with the best balance between robustness and speed. The model achieved scores of 0.43 MI_DSC, 0.47 MI_NSD, and runs at 69 FPS. For comparison, this model outperforms team *www*'s by 11.5% on MI_DSC and 12.1% on MI_NSD while running 13.8× faster.

Our most robust network and proposed architecture, namely *CBAM-Full + Aug + Anch + MS*, reached 13.7% MI_DSC and 13.9% MI_NSD scores higher, compared to the top contestant of the challenge (improvement over 44% and 39% on MI_DSC and MI_NSD, respectively). It also outperforms *CBAM-Full + Aug + Anch* with metric values showing 2.2% and 1.8% higher on MI_DSC and MI_NSD, respectively. However, *CBAM-Full + Aug + Anch + MS*'s increased complexity has an impact on its inference speed yielding 24 FPS. Nevertheless, the model is still 4.8× faster than the previous state-of-the-art. Table 4 shows the results for each development stage of our final network on all three test datasets. It can be observed that for other Stage 1 and Stage 2 as well our proposed architecture with CBAM-Full + Aug + Anch showed improved performance over most combinations. However, in these cases the addition of the multi-scale feature fusion network (MS) only provided competitive result (e.g., 0.46 and 0.43 on MI_DSC for Stage 1 and Stage 2, respectively).

Fig. 4 shows the dot-and-boxplots of the MI_DSC and MI_NSD metric values obtained by each of our algorithms on Stage 3 test set used in the challenge. We can observe a large difference between the top model and the baseline model, as well as the progressive improvement from experiment to experiment. Despite of the fact that most of these models are similar in terms of their median, the improvement is evident when looking at the aggregated metric values, as well as the first and third quartiles, with our final model having the least deviations.

4.5.2. Qualitative results

To better understand the effects of different components of our models, we performed a comparative analysis of frames with the worst performance of each network, as show in Fig. 5. Fig. 5(a) shows the qualitative comparison between worst frames of *Base YOLACT++* and their corresponding frames from *CBAM-Full*. We can observe from the images on the left side that the baseline model often presents missed detections, which hinders the model's recall. Similarly, the model confuses tissue and other objects like bandages as instruments, evidencing its lack of robustness. On the other hand, *CBAM-Full* overcomes some of these problems by attending to the important features, leading to

Table 3
Evaluation results for stage 3 of the challenge.

Team/Model	Base method	Aggr. MI_DSC	Aggr. MI_NSD	FPS
www	Mask R-CNN	0.31	0.35	5*
Uniandes	Mask R-CNN	0.26	0.29	5*
SQUASH	Mask R-CNN	0.22	0.26	5*
CASIA_SRL	Attention Network	0.19	0.27	5*
fisensee	2D U-Net	0.17	0.16	12*
VIE	Mask R-CNN	0.00	0.00	5*
Base YOLACT++	YOLACT++	0.00	0.00	75
CCAM-FPN	YOLACT++	0.00	0.00	60
CBAM-Backbone	YOLACT++	0.25	0.29	65
CCAM-Full	YOLACT++	0.30	0.30	45
CCAM-Backbone	YOLACT++	0.31	0.33	49
CBAM-FPN	YOLACT++	0.31	0.33	66
CBAM-Full	YOLACT++	0.34	0.38	65
CBAM-Full + Aug	YOLACT++	0.38	0.43	63
CBAM-Full + Aug + Anch	YOLACT++	0.43	0.47	**69**
CBAM-Full + Aug + Anch + MS	YOLACT++	**0.45**	**0.49**	24

The upper part of the table shows the aggregated metrics for the competitors of the 2019 ROBUST-MIS challenge. The lower part of the table shows the results of the our developed models. MI_DSC and MI_NSD metrics are reported along with the base models and frame rates.
*Approximated from base method. Original measurement was not reported.

Table 4
Metric comparison for all three test stages of ROBUST-MIS dataset.

Model	Stage 1		Stage 2		Stage 3	
	MI_DSC	MI_NSD	MI_DSC	MI_NSD	MI_DSC	MI_NSD
Base YOLACT++	0.31	0.33	0.00	0.00	0.00	0.00
CCAM-Backbone	0.4	0.47	0.32	0.33	0.31	0.33
CCAM-FPN	0.23	0.33	0.00	0.00	0.00	0.00
CCAM-Full	0.40	0.45	0.28	0.33	0.31	0.33
CBAM-Backbone	0.43	0.46	0.32	0.35	0.25	0.29
CBAM-FPN	0.43	0.49	0.33	0.37	0.31	0.33
CBAM-Full	0.40	0.47	0.33	0.38	0.34	0.38
CBAM-Full + Aug	0.45	0.5	0.31	0.38	0.38	0.43
CBAM-Full + Aug + Anch	0.46	**0.50**	**0.44**	0.47	0.43	0.47
CBAM-Full + Aug + Anch + MS	**0.46**	0.49	0.43	**0.48**	**0.45**	**0.49**

Best metric values are shown in bold for each test dataset.

Table 5
Evaluation results on the EndoVis 2017 challenge dataset for some recent methods and the nine configurations of our proposed model. All the performance (FPS) results for the Yolact model are given for 3 instances in the frame.

Model	Mean IoU	FPS
U-Net	56.87	12*
TernausNet	80.34	10*
RASNet	90.33	5*
Base YOLACT++	79.90	**71**
CCAM-Backbone	84.14	52
CCAM-FPN	84.80	62
CCAM-Full	84.82	52
CBAM-Backbone	85.00	62
CBAM-FPN	85.20	64
CBAM-Full	86.25	58
CBAM-Full + Aug	86.50	60
CBAM-Full + Aug + Anch	86.80	60
CBAM-Full + Aug + Anch + MS	**87.00**	25

Table 6
Execution times for stage 3 of the ROBUST-MIS challenge (in FPS).

Team/Model	1 inst.	2 inst.	3 inst.
Base YOLACT++	74	72	64
CCAM-Backbone	53	49	44
CCAM-FPN	60	58	60
CCAM-Full	50	47	52
CCAM-Backbone	65	63	55
CBAM-FPN	68	64	57
CBAM-Full	61	57	56
CBAM-Full + Aug	63	58	53
CBAM-Full + Aug + Anch	68	65	59
CBAM-Full + Aug + Anch + MS	26	24	25

Each of the models was tested with different sequences of videos, containing 1, 2 and 3 instrument instances respectively. Reported values were averaged over 50 runs.

better localization and segmentation results. We identified four different conditions that seemed particularly challenging for *CBAM-Full*: transparent instruments, vertical instruments, small instruments on the edge of the field of view, and partially occluded instruments. According to Roßet al. (2021), most of these issues have also been challenging to previous participants. Fig. 5(b) illustrates the comparison of worst frames of *CBAM-Full* and the improvements obtained after training with target-domain data augmentation.

In contrast to the baseline model with attention, we can observe that *CBAM-Full + Aug* is capable of addressing problematic instances such as small instruments on the edge of the field of view. Similarly, transparent, partially occluded, and vertical instruments are now detected and segmented to a larger extent.

Nonetheless, it must be emphasized that the *CBAM-Full + Aug* model still presented recurrent missed detections on long vertical and transparent instruments. For comparison, Fig. 5(c) illustrates its worst cases and the improvements obtained after training this model with an optimized set of anchors to combat this issue. As we can observe in the figure, the anchor optimization in the *CBAM-Full + Aug + Anch* model led to additional detection and segmentation improvements not only on previously undetected objects but across all instruments.

On the other hand Fig. 5(d) illustrates the comparison of worst frames of *CBAM-Full + Aug + Anch* and the improvements obtained after training with MSFF. We can observe that *CBAM-Full + Aug + Anch*

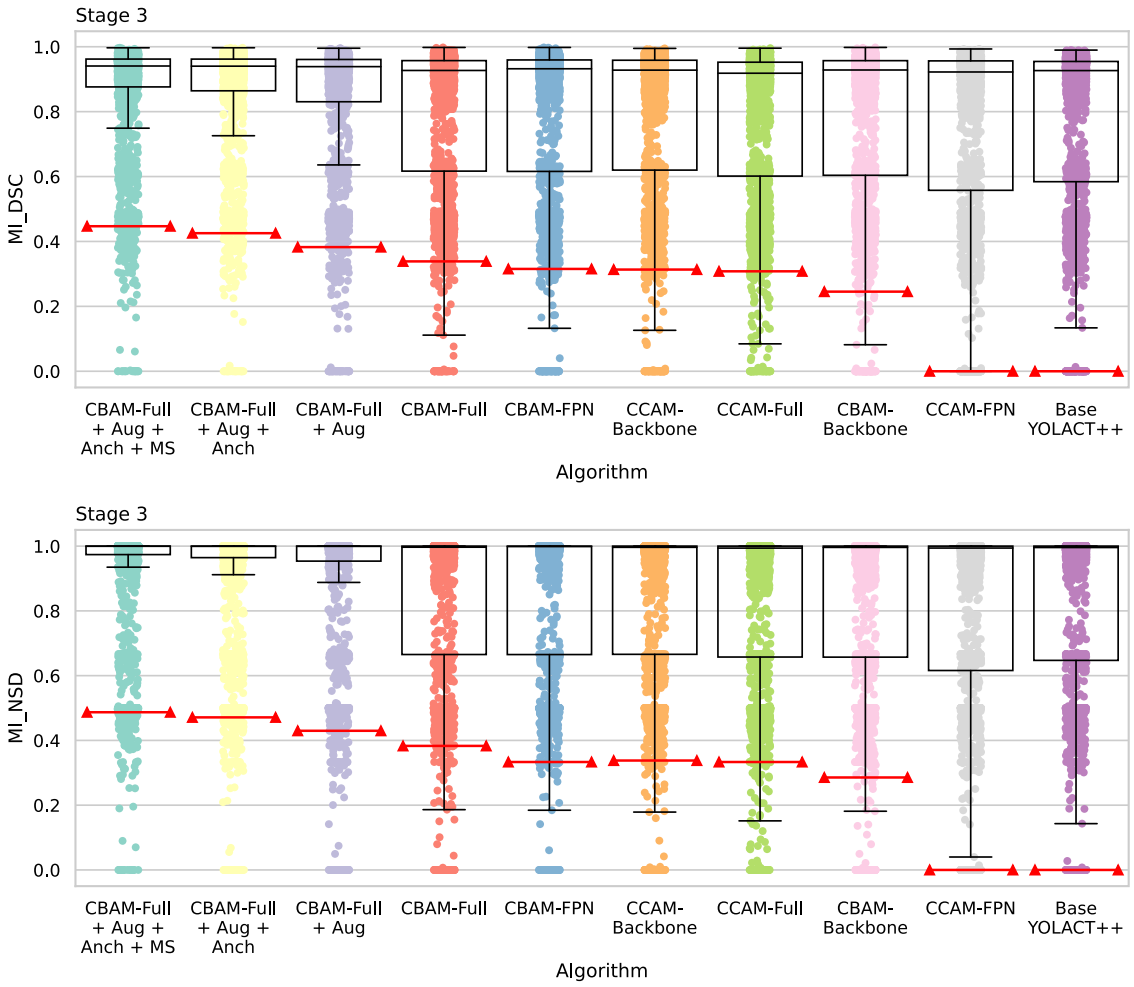


Fig. 4. Dot-and-boxplots for the MI_DSC and MI_NSD. Individual performances of algorithms on stage 3 of the challenge are shown. The red lines indicate the value of the aggregated metric (by 5% percentile) for each algorithm.

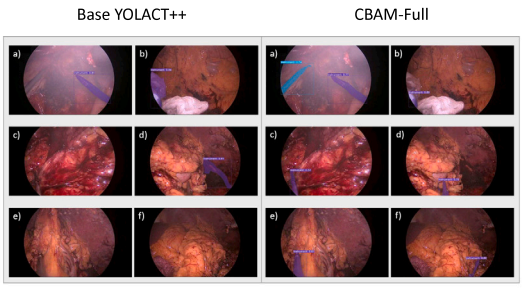
+ *MS* is better at recognizing more challenging instances, like small under-exposed instruments, transparent instruments, and reflections.

4.6. Results on the EndoVis robotic instrument segmentation 2017 dataset

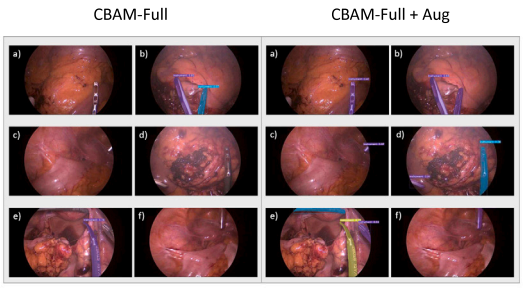
In order to assess the applicability of our model beyond the Robust MIS dataset, we carried out experiments on previous MICCAI EndoVis Robotic Instrument Segmentation Challenge 2017 dataset (Shvets et al., 2018). We trained all nine different configurations of our Yolact++-based architecture (described in Section 3.2). The Robotic Instrument Segmentation challenge consisted of three sub-tasks: a) binary instrument segmentation, b) instrument part segmentation, and c) instrument type segmentation tasks. Our experiments reported in this work is aimed at the instrument type segmentation task as it is similar to the ROBUST-MIS challenge.

The EndoVis-2017 Robotic Instrument Segmentation Challenge dataset consists of 10 sequences of abdominal porcine procedures recorded using the da Vinci Xi robotic system (Shvets et al., 2018). The Surgical instruments were divided into six categories, namely Bipolar Forceps, Prograsp Forceps, Needle Driver, Vessel Sealer, Grasping Retractor and Curved Scissors. A miscellaneous category is labeled for any other surgical instrument. For the original challenge, the organizers provided the first 225 frames of 8 sequences as training data and kept the last 75 frames of those sequences as test data. Additionally, two of the full 300 frame sequences were kept as test sequences. The final dataset is comprised of 1800 images with a resolution of 1920 × 1080 (1400 for training and 400 for test). Since the labels of test set are not available for model evaluation, we followed the same protocol for training and testing as reported in the Refined Attention Segmentation Network (RASNet) (Ni et al., 2019). We selected consecutive sequences to avoid similar frames in the training and test set. We kept the 1350 images as the training set and reserved 450 images for the test set. We did not use any additional data for training or fine-tuning our nine model configurations and the baseline methods.

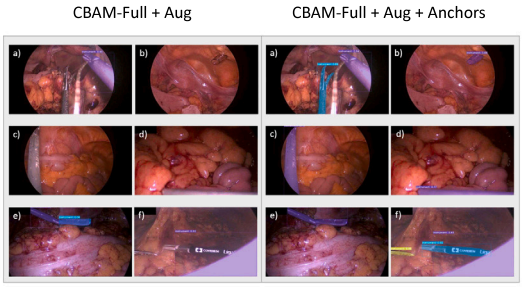
The results are summarized in Table 5. We compare our results with three baseline methods used in Ni et al. (2019): (i) a simple U-Net architecture (Ronneberger et al., 2015a), (ii) TernausNet (Iglovikov and Shvets, 2021), a U-Net-like architecture that uses a VGG16 backbone as an encoder module and (iii) RASNet (Ni et al., 2019), which makes use of an encoder–decoder structure extended with an attention fusion modules (AFM) to combine multi-scale features in a similar manner to our work. From the table, it can be observed that our method outperforms other methods originally presented in the EndoVis Robotic Instrument Segmentation Challenge (UNet and TernausNet) and falls slightly behind RASNet (87% vs 90% mIoU). Nonetheless, all these previous methods run at much slower frame rates than ours. In fact, as shown in the last column of table, our base model runs at 71 FPS whereas our proposed full model configuration can attain 25 FPS (for 3 instances) which is still 5 times faster than RASNet. An optimal choice between speed and accuracy would be our CBAM-Full+Aug+Anch model which is 12 times faster and with an mean IoU of 86.80.



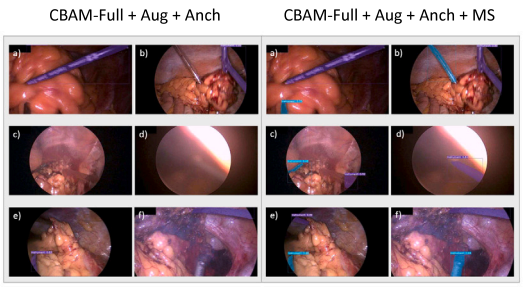
(a) Qualitative improvements of worst performing frames in *Base YOLACT++* that are handled by our *CBAM-Full*. Failure cases for base model are: a) Missed instrument detection due to smoke overlay, b) false positive segmentation of bandage as instrument, c) missed detection due to blood overlay, d) false positives in segmentation of blood pool and shadow, e) missed detection due to motion blur, and f) no detection of instrument that is covered in tissue.



(b) Qualitative improvements of worst performing frames in *CBAM-Full* that are handled by our *CBAM-Full + Aug*. Failure cases for *CBAM-Full* are: a) Missed vertical instrument, b) single (split) instrument mistakenly regarded as two instances, c) small instrument on the edge of the field of view, d) missed vertical instrument and missed instrument on the edge of the field of view, e) missed instruments on the edge of FOV, and f) missed detection of a transparent instrument.



(c) Qualitative improvements of worst performing frames in *CBAM-Full + Aug* that are handled by our *CBAM-Full + Aug + Anch*. Failure cases for *CBAM-Full + Aug* are: a) Missed vertical instrument, b) missed transparent instrument, c) missed vertical instrument on the edge of the field of view, d) missed long horizontal instrument on the bottom of the image, e) single instrument mistaken as two instruments and overlapped masks, and f) missed detection of two long and horizontal instruments.



(d) Qualitative improvements of worst performing frames in *CBAM-Full + Aug + Anch* that are handled by our *CBAM-Full + Aug + Anch + MS*. Failure cases for *CBAM-Full + Aug + Anch* are: a) Missed underexposed instrument, b) missed transparent instrument, c) missed blurred instruments, d) missed blurred instrument due to reflection, e) missed underexposed instrument, and f) missed detection due to motion artifacts.

Fig. 5. Qualitative comparison of challenging cases showing incremental improvements from baseline to our proposed final model.

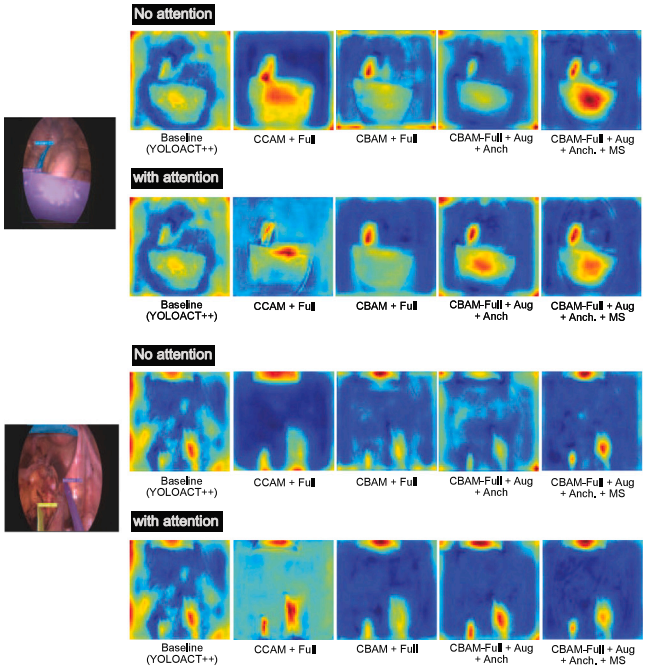


Fig. 6. Network output with and without attention. (top) Large part of image covered with instrument in purple and in blue (left). (bottom) Three instruments two at the bottom of image and one on the top of image (see left). In visual saliency maps, the dark blue color indicates lowest activation values, while bright red indicates higher values.

5. Discussion

While deep learning has allowed us to design data driven solutions, generalization remains a key issue that can cause significant performance degradation, especially in instrument segmentation applications where the intervention objects are of variable shapes and exposed to dynamic environments (e.g., smoke, flares, specularity etc, see Fig. 1). In surgery, real-time performance of such tool is of enormous importance for clinical utility. However, most methods built in the past relied on two stage networks that are not computationally efficient.

In order to tackle previous methods limitations, we built over single stage YOLACT (Bolya et al., 2019) and added several modifications to improve both the accuracy and robustness of our final surgical instrument segmentation algorithm. Among our models, we observed that those based on CBAM achieved slightly better performance that the ones based on CCAM. Regardless, attention-integrated models always outperformed the attention-less baseline in terms of robustness. Fig. 4 shows dot-and-boxplots of the metric values obtained by each algorithm over all test cases in stage 3 dataset (unseen test dataset in the ROBUST-MIS challenge). We can observe that adding attention mechanisms boosts the performance compared to the baseline model used in our architecture design (Bolya et al., 2019). This especially true for instances below the third quartile, which are the most important for our performance metrics. Among the three model variations to which we added CCAM attention modules, CCAM-Backbone achieved the best results in terms of robustness (0.313 MI_DSC and 0.338 MI_NSD, see Table 3). This indicates that the contextually enriched feature maps from the ResNet-101 backbone are powerful enough to generate more accurate mask prototypes and coefficients in the YOLACT architecture better segmentation outputs.

However, it must be noted that the CBAM-Full model outperformed all other attention modules by nearly 9% on MI_DSC over the best performing CCAM-Backbone. The superiority of CBAM when integrated together with backbone and FPN all together enhances both channel-wise and block attentions better to represent local and global features well. The FPN layer allows to capture size variability present in the dataset for which CCAM yielded zero on aggregated MI_DSC score (Fig. 4).

We thoroughly investigated on model improvement for generalizability through different mechanisms such as data augmentation and optimization of anchor weights. These steps were experimentally proven to be right directions giving subsequent increase in both aggregated MI_DSC and MI_NSD (see Table 3, Figs. 5(a) and 5(b)). For example, by using augmentation, we observed that the images with different view points (mostly long sized instruments either straight or slightly oblique) performed better; this can be because of low number of such samples in the dataset (Fig. 5(b)). Similarly, for the variable size instruments optimizing anchors provided substantial improvements (Fig. 5(c)). We also noticed issues with the small instrument appearing either on the bottom or sides of the frame, and also for those which appeared as background, due to specularity or covering large tissue area with similar color (see Fig. 5(c), left). The use of the Multi-scale feature fusion (MSFF) network allowed us to fuse features at different scales and layers (Fig. 5(d)). Our architecture with this MSFF network integrated together with the attention maps allowed to transfer both local and global context fusing high and low-level feature representations.

As a result of integrating MSFF, we observed further 2% higher performance on aggregated MI_DSC above the CBAM-Full model (Table 3). Our final best approach has near real-time performance of 24 FPS on NVIDIA Volta GPU. However, our second best performing method can run at 69 FPS which is above the required real-time performance in most cases (i.e., 45 FPS). A similar, performance improvement was observed for other two test datasets (Table 4).

To better understand the behavior of the two types of attention modules used in our experiments, we visualized the attention maps generated by 4 different models, in addition to the feature maps of the baseline model (see Fig. 6). We chose the *CBAM-Full*, *CCAM-Full*, and *CBAM-Full + Aug + Anch + MS* configurations for visualization purposes as they show the refined feature maps using 2 types of attention modules. Fig. 6 illustrates the activation maps taken before and after the attention modules. We are interested in the effects of attention on this specific point of the network since mask prototypes and coefficients are extracted from FPN features and backbone network. We can observe that over-mixing problem of *CCAM-Full*. In fact, the maps before CCAM in the Full model are cleaner and more discriminative than the actual attention output.

This result confirms that CCAM excels at refining features that have not been mixed before. On the other hand, CBAM-based models produce slightly lower quality feature maps from attention modules in the backbone, as it can be observed by the clouds and blobs on the top maps of the figure. Nonetheless, the features are drastically refined after passing the attention modules in the Full model configuration, leading to clean and discriminative activations. This corroborates our hypothesis that CBAM is better at refining previously aggregated data, yielding a superior overall performance.

Nonetheless, as shown in the last column of Table 3, these improvements in terms of robustness come at the price of an increased inference time, impacting the attainable real-time performance of our model. In order to further understand the effect the various optimizations of the model, as well as the impact of the number of instrument instances present in the frame on the model's performance, we carried out various inference speed assessments by running inference tests on three different 10-second video snippets from the ROBUST-MIS dataset. The results of this experiment are summarized in Table 6. The first video contains only one instrument instance, the second contains two instruments and the third one contains three instrument instances in the same frame. Each video was evaluated a total of fifty times per model, and the reported frame rates were then aggregated by the mean. Inference time was tested on a single Tesla P100 GPU using a batch size of 1. The results of this experiment show that there is indeed a loss in performance as the number of instances within the frame is increased; in average, there is a decrease of about 2 to 4 FPS when the number of instances is increased to two and of 10 FPS when 3 instances are in the frame. As discussed above, our best model is able to attain very high scores with a performance of about 24 FPS; it must be noted that this model is more robust to changes in the number of instances in the frame, as it can be observed in the last row of Table 6.

Several interesting findings aroused from these experiments. For instance, as it can be observed in the table, there is a noticeable difference in inference time between the CBAM-Full and CBAM-Full + Aug models, which might be surprising at first glance as they are in fact the same model. However, this gap can be due to the difference in the number of instances predicted during inference, which would vary with the degree of domain-targeted data augmentation used for training the model. In our architecture (see Fig. 2), the prediction head is connected to subsequent post processing steps such as NMS, region cropping and thresholding. Ideally, fewer false positives should yield a higher FPS. However, the FPS difference of 2 between CBAM-Full and CBAM-Full+Aug could be due to outlier samples. Similarly, if one looks at the CBAM-Full+Aug+Anchor which improves the MI_DSC by 10%, while providing a better FPS, as it minimizes the false positives.

Furthermore, domain-targeted data augmentation improves the accuracy of the model thereby creating less false positives during inference. This reduces the computational overhead introduced by the prediction head. As a result, other post processing steps such as NMS, region cropping and thresholding consume less time compared to the CBAM-Full model, which can bee seen as sub-optimally trained model

6. Conclusion

We have developed a real-time novel architecture that builds upon a single-stage instance segmentation method. Through our step-wise solution to different problems in surgical data, we have identified and integrated components that can be deal with existing and eminent challenges for robust segmentation of surgical instruments. We have provided comprehensive experiments and analysis that supports our final architecture development and its impact on surgical tool segmentation in clinic. Our method outperformed all methods reported in recently conducted ROBUST-MIS challenge.

Our current work is built upon publicly available retrospective dataset providing a strong evidence of robustness ability compared to presented methods at the ROBUST-MIS challenge and current literature. In future work, we will validate on prospective data and benchmark it in clinical settings.

CRediT authorship contribution statement

Juan Carlos Ángeles Cerón: Conceptualization, Methodology, Investigation, Software, Data curation, Writing – original draft; Writing – review & editing, Visualisation. **Gilberto Ochoa Ruiz:** Conceptualisation, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision. **Leonardo Chang:** Conceptualisation, Supervision. **Sharib Ali:** Conceptualisation, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the AI Hub and the Centro de Internet de las Cosas (CIOT) at Tecnologico de Monterrey for their support for carrying the experiments reported in this paper on their NVIDIA's DGX computer.

References

Ali, S., Dmitrieva, M., Ghatwary, N., Bano, S., Polat, G., Temizel, A., Krenzer, A., Hekalo, A., Guo, Y.B., Matuszewski, B., Gridach, M., Voiculescu, I., Yoganand, V., Chavan, A., Raj, A., Nguyen, N.T., Tran, D.Q., Huynh, L.D., Boutry, N., Rezvy, S., Chen, H., Choi, Y.H., Subramanian, A., Balasubramanian, V., Gao, X.W., Hu, H., Liao, Y., Stoyanov, D., Daul, C., Realdon, S., Cannizzaro, R., Lamarque, D., Tran-Nguyen, T., Bailey, A., Braden, B., East, J.E., Rittscher, J., 2021. Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy. Med. Image Anal. 70, 102002. http://dx.doi.org/10.1016/j.media.2021.102002.

Ali, S., Zhou, F., Braden, B., Bailey, A., Yang, S., Cheng, G., Zhang, P., Li, X., Kayser, M., Soberanis-Mukul, R.D., Albarqouni, S., Wang, X., Wang, C., Watanabe, S., Oksuz, I., Ning, Q., Yang, S., Khan, M.A., Gao, X.W., Realdon, S., Loshchenov, M., Schnabel, J.A., East, J.E., Wagnieres, G., Loschenov, V.B., Grisan, E., Daul, C., Blondel, W., Rittscher, J., 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci. Rep. 10, 2748. http://dx.doi.org/10.1038/s41598-020-59413-5.

Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes-Hurtado, F., Flouty, E., Mohammed, A.K., Pedersen, M., Kori, A., Varghese, A., Krishnamurthi, G., Rauber, D., Mendel, R., Palm, C., Bano, S., Saibro, G., Shih, C.S., Chiang, H.A., Zhuang, J., Yang, J., Iglovikov, V.I., Dobrenkii, A., Reddiboina, M., Reddy, A., Liu, X., Gao, C., Unberath, M., Azizian, M., Stoyanov, D., Maier-Hein, L., Speidel, S., 2020. 2018 Robotic scene segmentation challenge. ArXiv abs/2001.11190.

Allan, M., Shvets, A.A., Kurmann, T., Zhang, Z.V., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., García-Peraza, W., Iglovikov, V.I., Luo, H., Yang, J., Stoyanov, D., Maier-Hein, L., Speidel, S., Azizian, M., 2019. 2017 Robotic instrument segmentation challenge. ArXiv abs/1902.06426.

Bartoli, A., Collins, T., Bourdel, N., Canis, M., 2012. Computer assisted minimally invasive surgery: is medical computer vision the answer to improving laparosurgery? Med. Hypotheses 79, 858–863.

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., Speidel, S., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv:1805.02475.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2019. YOLACT: real-time instance segmentation. CoRR abs/1904.02689.

Bolya, D., Zhou, C., Xiao, F., Lee, Y.J., 2020. YOLACT++: better real-time instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 1. http://dx.doi.org/10.1109/tpami.2020.3014297.

Caroff, D.A., Chan, C., Kleinman, K., Calderwood, M.S., Wolf, R., Wick, E.C., Platt, R., Huang, S., 2019. Association of open approach vs laparoscopic approach with risk of surgical site infection after colon surgery. JAMA Netw. Open 2, e1913570.

Chaudhari, S., Polatkan, G., Ramanath, R., Mithal, V., 2019. An attentive survey of attention models. ArXiv abs/1904.02874.

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y., 2020. Blendmask: Top-down meets bottom-up for instance segmentation. arXiv:2001.00309.

Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder–decoder with atrous separable convolution for semantic image segmentation. CoRR abs/1802.02611.

Ding, H., Jiang, X., Shuai, B., Liu, A.Q., Wang, G., 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2393–2402. http://dx.doi.org/10.1109/CVPR.2018.00254.

Fu, Z., Jin, Z., Zhang, C., He, Z., Zha, Z., Hu, C., Gan, T., Yan, Q., Wang, P., Ye, X., 2021. The future of endoscopic navigation: A review of advanced endoscopic vision technology. IEEE Access 9, 41144–41167.

Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., Zhang, S., 2020. Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/tmi.2020.3035253.

He, K., Gkioxari, G., Dollár, P., Girshick, R.B., 2017. Mask R-CNN. CoRR abs/1703.06870. arXiv:1703.06870.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2018. Mask r-cnn. arXiv:1703.06870.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv:1512.03385.

Huang, Z., Wang, X., Wei, Y., Huang, L., Shi, H., Liu, W., Huang, T.S., 2020. Ccnet: Criss-cross attention for semantic segmentation. arXiv:1811.11721.

Hurtik, P., Molek, V., Hula, J., Vajgl, M., Vlasanek, P., Nejezchleba, T., 2020. Poly-yolo: higher speed, more precise detection and instance segmentation for yolov3. arXiv:2005.13243.

Iglovikov, V.I., Shvets, A.A., 2021. TernausNet. Springer International Publishing, Cham, pp. 127–132. http://dx.doi.org/10.1007/978-3-030-64340-9_15.

Isensee, F., Maier-Hein, K.H., 2020. Or-unet: an optimized robust residual u-net for instrument segmentation in endoscopic images. arXiv:2004.12668.

Jha, D., Ali, S., Tomar, N.K., Riegler, M.A., Johansen, D., Johansen, H.D., Halvorsen, P., 2021. Exploring deep learning methods for real-time surgical instrument segmentation in laparoscopy. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics. BHI, pp. 1–4. http://dx.doi.org/10.1109/BHI50953.2021.9508610.

Kaul, C., Manandhar, S., Pears, N., 2019. Focusnet: An attention-based fully convolutional network for medical image segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, pp. 455–458. http://dx.doi.org/10.1109/ISBI.2019.8759477.

Lee, Y., Park, J., 2019. Centermask : Real-time anchor-free instance segmentation. CoRR abs/1911.06667.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017c. Feature pyramid networks for object detection. arXiv:1612.03144.

Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P., 2017b. Focal loss for dense object detection. CoRR abs/1708.02002.

Lin, G., Milan, A., Shen, C., Reid, I., 2017a. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1925–1934.

Long, J., Shelhamer, E., Darrell, T., 2015a. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3431–3440. http://dx.doi.org/10.1109/CVPR.2015.7298965.

Long, J., Shelhamer, E., Darrell, T., 2015b. Fully convolutional networks for semantic segmentation. arXiv:1411.4038.

Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., Nakawala, H., Park, A., Pugh, C., Stoyanov, D., Vedula, S.S., Cleary, K., Fichtinger, G., Forestier, G., Gibaud, B., Grantcharov, T., Hashizume, M., Heckmann-Nötzel, D., Kenngott, H.G., Kikinis, R., Mündermann, N., Onogur, S., Sznitman, R., Taylor, R.H., Tizabi, M.D., Wagner, M., Hager, G.D., Neumuth, T., Padoy, N., Collins, J., Gockel, I., Goedeke, J., Hashimoto, D.A., Joyeux, L., Lam, K., Leff, D.R., Madani, A., Marcus, H.J., Meireles, O., Seitel, A., Teber, D., Ückert, F., Müller-Stich, P., Speidel, S., 2021a. Surgical data science – from concepts toward clinical translation. arXiv:2011.02284.

Maier-Hein, L., Wagner, M., Ross, T., Reinke, A., Bodenstedt, S., Full, P.M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Kisilenko, A., Müller, B., Davitashvili, T., Capek, M., Tizabi, M.D., Eisenmann, M., Adler, T.J., Gröhl, J., Schellenberg, M., Seidlitz, S., Lai, T.Y.E., Pekdemir, B., Roethlingshoefer, V., Both, F., Bittel, S., Mengler, M., Mündermann, L., Apitz, M., Kopp-Schneider, A., Speidel, S., Nickel, F., Probst, P., Kenngott, H.G., Müller-Stich, B.P., 2021b. Heidelberg colorectal data set for surgical data science in the sensor operating room. Sci. Data 8, 101. http://dx.doi.org/10.1038/s41597-021-00882-2.

Milletari, F., Rieke, N., Baust, M., Esposito, M., Navab, N., 2018. Cfcm: Segmentation via coarse to fine context memory. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2018, Springer International Publishing, Cham, pp. 667–674.

Ni, Z.L., Bian, G.B., Xie, X.L., Hou, Z.G., Zhou, X.H., Zhou, Y.J., 2019. Rasnet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBC, pp. 5735–5738. http://dx.doi.org/10.1109/EMBC.2019.8856495.

Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. CoRR abs/1506.01497.

Ronneberger, O., Fischer, P., Brox, T., 2015a. U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention. MICCAI 2015, Springer International Publishing, Cham, pp. 234–241.

Ronneberger, O., Fischer, P., Brox, T., 2015b. U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597. arXiv:1505.04597.

Roß, T., Reinke, A., 2019. Robustmis2019. URL: https://phabricator.mitk.org/source/rmis2019/.

Roß, T., Reinke, A., Full, P.M., Wagner, M., Kenngott, H., Apitz, M., Hempe, H., Mindroc-Filimon, D., Scholz, P., Tran, T.N., Bruno, P., Arbeláez, P., Bian, G.B., Bodenstedt, S., Bolmgren, J.L., Bravo-Sánchez, L., Chen, H.B., González, C., Guo, D., Halvorsen, P., Heng, P.A., Hosgor, E., Hou, Z.G., Isensee, F., Jha, D., Jiang, T., Jin, Y., Kirtac, K., Kletz, S., Leger, S., Li, Z., Maier-Hein, K.H., Ni, Z.L., Riegler, M.A., Schoeffmann, K., Shi, R., Speidel, S., Stenzel, M., Twick, I., Wang, G., Wang, J., Wang, L., Wang, L., Zhang, Y., Zhou, Y.J., Zhu, L., Wiesenfarth, M., Kopp-Schneider, A., Müller-Stich, L., 2021. Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the robust-mis 2019 challenge. Med. Image Anal. 70, 101920. http://dx.doi.org/10.1016/j.media.2020.101920.

Seo, P.H., Lin, Z., Cohen, S., Shen, X., Han, B., 2018. Progressive attention networks for visual attribute prediction. arXiv:1606.02393.

Sheetz, K.H., Claflin, J., Dimick, J.B., 2020. Trends in the adoption of robotic surgery for common surgical procedures. JAMA Netw. Open 3, e1918911.

Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications. ICMLA, pp. 624–628. http://dx.doi.org/10.1109/ICMLA.2018.00100.

Sinha, A., Dolz, J., 2021. Multi-scale self-guided attention for medical image segmentation. IEEE J. Biomed. Health Inf. 25, 121–130.

Wang, Y., Deng, Z., Hu, X., Zhu, L., Yang, X., Xu, X., Heng, P.A., Ni, D., 2018. Deep attentional features for prostate segmentation in ultrasound. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention. MICCAI 2018, Springer International Publishing, Cham, pp. 523–530.

Wang, J., Sun, K., Cheng, T.., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B., 2020a. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1. http://dx.doi.org/10.1109/TPAMI.2020.2983686.

Wang, X., Zhang, R., Kong, T., Li, L., Shen, C., 2020b. Solov2: Dynamic and fast instance segmentation. arXiv:2003.10152.

Ward, T.M., Mascagni, P., Ban, Y., Rosman, G., Padoy, N., Meireles, O., Hashimoto, D.A., 2021. Computer vision in surgery. Surgery 169, 1253–1256. http://dx.doi.org/10.1016/j.surg.2020.10.039.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11, 1–15.

Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. arXiv:1807.06521.

Xu, Y., Li, Y., Liu, M., Wang, Y., Lai, M., Chang, E.I.C., 2016. Gland instance segmentation by deep multichannel side supervision. arXiv:1607.03222.

Xu, J., Wang, W., Wang, H., Guo, J., 2020. Multi-model ensemble with rich spatial information for object detection. Pattern Recognit. 99, 107098. http://dx.doi.org/10.1016/j.patcog.2019.107098.

You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 4651–4659. http://dx.doi.org/10.1109/CVPR.2016.503.

Zhao, Z., Jin, Y., Chen, J., Lu, B., Ng, C.F., Liu, Y.H., Dou, Q., Heng, P.A., 2021. Anchor-guided online meta adaptation for fast one-shot instrument segmentation from robotic surgical videos. Med. Image Anal. 74, 102240. http://dx.doi.org/10.1016/j.media.2021.102240.

Zlocha, M., Dou, Q., Glocker, B., 2019. Improving retinanet for ct lesion detection with dense masks from weak recist labels. arXiv:1906.02283.