Endometriosis detection and localization in laparoscopic gynecology

Andreas Leibetseder¹ · Klaus Schoeffmann¹ · Jörg Keckstein² · Simon Keckstein³

Received: 17 March 2021 / Revised: 30 September 2021 / Accepted: 2 November 2021 /
Published online: 7 January 2022
© The Author(s) 2021

Abstract

Endometriosis is a common gynecologic condition typically treated via laparoscopic surgery. Its visual versatility makes it hard to identify for non-specialized physicians and challenging to classify or localize via computer-aided analysis. In this work, we take a first step in the direction of localized endometriosis recognition in laparoscopic gynecology videos using region-based deep neural networks Faster R-CNN and Mask R-CNN. We in particular use and further develop publicly available data for transfer learning deep detection models according to distinctive visual lesion characteristics. Subsequently, we evaluate the performance impact of different data augmentation techniques, including selected geometrical and visual transformations, specular reflection removal as well as region tracking across video frames. Finally, particular attention is given to creating reasonable data segmentation for training, validation and testing. The best performing result surprisingly is achieved by randomly applying simple cropping combined with rotation, resulting in a mean average segmentation precision of 32.4% at 50-95% intersection over union overlap (64.2% for 50% overlap).

Keywords Endometriosis segmentation · Lesion detection · Deep learning

✉ Andreas Leibetseder
aleibets@itec.aau.at

Klaus Schoeffmann
ks@itec.aau.at

Jörg Keckstein
joerg@keckstein.at

Simon Keckstein
simon.keckstein@med.uni-muenchen.de

¹ Institute of Information Technology, Klagenfurt University, Klagenfurt, Austria

² Medical Faculty, Ulm University, Ulm, Germany

³ University Hospital, Ludwig-Maximilians-University Munich, Munich, Germany

1 Introduction

Minimally invasive surgery (MIS) in gynecology focuses on the endoscopic treatment of conditions and diseases of the female inner genital tract. It is conducted by using special camera systems – *endoscopes* – as well as surgical instruments introduced through small incisions into the abdomen of a patient. In this way, the physician is able to perform treatments via observing the camera feed, while greatly reducing the physical injury imposed on patients as well as facilitating intervention monitoring, computer-aided live assistance and even the creation of medical video archives for post-surgical analysis. In particular, the latter benefit enables physicians to revisit interventions at any time for gaining useful insights or improving treatment planning and medical education.

Revisiting intervention archives can especially be important for the treatment of patients with *endometriosis* and infertility. These patients suffer from *endometrial*-like tissue in a variety of abnormal locations external to the uterus. In over 60% of diagnosed cases, endometriosis treatment requires multiple surgeries [34] due to unidentified (missed) lesions, which can be mitigated by consulting archived intervention videos. Depending on the practice of an individual physician, it certainly is possible to record specific short clips of relevant content during surgery, however, typically surgeries are archived in full meaning that the resulting videos generally contain a considerable amount of irrelevant content such as out-of-patient views, visually inexpressive or static scenes where treatment is interrupted and even common surgery phases that are of little importance for post-surgical analysis. Such properties combined with the typical routine of using conventional video players and file browsers for manually examining surgical video archives greatly impair revisiting past interventions, while also making this procedure error-prone. To improve this situation, computer-aided automatic content analysis can be employed for creating systems that highlight potentially relevant content to physicians during patient case inspections. However, although obvious irrelevant video segments such as overly blurry frames or camera testing screens can reasonably be identified via video analysis [24], more sophisticated approaches are required for identifying very specific content such as scenes showing endometriosis lesions.

The main symptoms of endometriosis, a benign inflammatory disease, are menstrual-related pelvic pain and unexplainable infertility. It is found in a large variety of locations and treatment strategies depend on the symptoms, the extent and localization of the lesions. The best known classification systems for endometriosis are the revised American Society for Reproductive Medicine (*rASRM*) score [1] and the *Enzian* classification scheme [14, 15]. The *rASRM* score describes superficial lesions of the peritoneum and ovarian endometriosis in four stages, whereas the *Enzian* classification categorizes deep endometriosis. Alongside these different possible locations, endometriotic lesions also strongly vary in their visual appearance, both intra- and interpersonal. This is illustrated by Fig. 1 showing a set of positive and negative endometriosis examples¹ at three common locations: ovary, uterus and peritoneum. In direct comparison and without a specific medical background, the differences between normal and pathological tissue are very difficult to discern, which evidently holds true for laymen but even inexperienced medical practitioners. Consequently, with the current successful application of deep

¹ Please note that Fig. 1b, d and f merely depict a subset of all the conditions' visual appearances – refer to Section 3.1 for more examples.

Fig. 1 Endometriosis examples – negative (left column) and positive (right column) examples at locations ovary, uterus and peritoneum

learning in many medical fields, attempting to solve this problem via computer-aided analysis seems reasonable. Moreover, being able to classify and potentially locate endometriosis can not only be helpful during interventions but also in treatment planning and particularly in teaching/training.

In this work we create, utilize and evaluate different datasets for a binary endometriosis detection and localization task using state-of-the-art region-based convolutional neural networks (R-CNNs). Since as of yet to the best of our knowledge no comparable work exists, this paper's findings can be considered benchmarks for future research on the topic. We define our contributions as follows:

Dataset development, Section 3.1. By revising the labeling strategy of publicly available endometriosis dataset GLENDa [20] towards visual similarity, we discover a large improvement in lesion segmentation performance. Further refinement of a newly

defined specifically well performing class – *endometrial implants* – leads to the release of a novel dataset named ENdometrial Implants Dataset² (ENID).

Data augmentation, Section 3.2. Utilizing ENID, we investigate the segmentation performance impact of a variety of data augmentation techniques including rotation, blurring, cropping, perspective transformation, reflection removal and frame-by-frame object tracking.

Performance evaluation, Sections 3.3& 3.4. For training deep models, we employ state-of-the-art object detection and segmentation models Faster R-CNN [33] and Mask R-CNN [9] using ResNet [8] variants as backbone networks. For creating training, validation and testing splits we divide the data twofold and pay specific attention to avoiding similar images across splits, thus risking overfitting our models.

This work is mainly partitioned into three sections: related work in Section 2, methodology description in Section 3 and discussion in Section 4.

2 Related work

The concept of learning features rather than defining them by hand had a tremendous impact on image classification over the last years. Deep convolutional networks such as AlexNet [17], GoogLeNet [35] or ResNet [8] have been successfully applied to countless domains, even non-image-related ones. Such networks not only are utilized on their own, they also represent valuable backbones to many deep architectures for image as well as video analysis. One such architecture family heavily using CNNs as backbones for region of interest (ROI) prediction and labeling are named region based convolutional neural networks or R-CNNs. These R-CNN networks after sufficient training are capable of locating, classifying and even segmenting objects in images through smart arrangement and utilization of CNN components. When regarding the evolution of R-CNNs from Fast R-CNN [6] over Faster R-CNN [33] up to Mask R-CNN [9], large improvements in runtime performance can be observed, as architecture bottlenecks were gradually resolved. Albeit said R-CNNs are developed for the common objects in context (COCO) challenge [21], which addresses the problem of real-world object detection, their successful application in many domains including medical imaging encourages us to as well employ variants of Faster R-CNN and Mask R-CNN architectures for this study's challenging lesion detection tasks.

The rising performance improvements of deep learning in medicine entails an ever-expanding range of applications targeted at providing medical staff with digital assistance for patient treatment. While similarly to the general multimedia domain, a large variety of data is analyzed ranging from electronic health records (E.H.R) over biosignals up to images produced by manifold clinical imaging technologies [30]. Apart from being used across many special fields, medical imaging also varies strongly depending on its purpose: monochrome images taken from computer tomography (CT) scans or ultrasound are very different compared to open surgery or endoscopic recordings. This in general renders scientific research on medical image classification more difficult to compare than in traditional multimedia analysis. Surprisingly, endoscopic images as of yet are not nearly as much digitally analyzed as many other technologies such as magnetic resonance imaging (MRI) [31,

² <http://ftp.itec.aau.at/datasets/ENID> or <https://doi.org/10.5281/zenodo.4570969>

32] or even the very similar field of microscopy [22]. Nevertheless, when regarding existing research on endoscopic video or image analysis three application areas can be defined: pre-processing methods, real-time support at procedure time and post-procedural applications [24]. Lesion detection in the laparoscopic field, as targeted in this study, can, depending on the use case, be placed into both of the latter categories. For instance, it can be helpful for tasks such as pointing out pathology during surgery as well as having post-surgical functions like automatically annotating medical video archives.

Generally, endoscopic surgery videos offer a great variety of research topics for applying deep learning. For instance, many studies are found on classifying or detecting content such as anatomy [19, 43], surgical tasks [29], instruments [12, 16, 26] or even the occurrence of smoke [18]. Furthermore, attempts have been made to recognize surgical phases based on that content [37, 42]. Finally, laparoscopic lesion segmentation using deep neural networks is a comparably scarcely researched field, which for a large part can be attributed to the limited availability of expert-annotated medical datasets.

While lesion detection is much addressed in the more restrictive field of gastrointestinal endoscopy [2], e.g. for polyp detection [11, 27, 28], to best of our knowledge, no directly comparable work has been conducted regarding image or video-based laparoscopic endometriosis segmentation. When, however, targeting classification we find almost all image-based works are on the different subject of endometrial cancer detection in MRI [38]. Nevertheless, we identify two approaches targeting endometriosis classification, albeit one does not contain any results besides extracting visual features [39]. The other approach uses the previously published GLENDa [20] dataset, which is as well used and further developed in this study, for binary classifying endometriotic regions using several CNNs [40]. They obtain classification accuracies between 80-90%. In the absence of any other related work to the best of our knowledge, we further attempt to give the reader an overview and list a small number of notable studies in the field of laparoscopic object detection and segmentation together with their findings. Predominantly, studies approaching object detection in laparoscopy attempt to localize instruments as is highlighted by the existence of literature surveys such as the one written by Yang et al. [41]. In fact, several studies achieve good results for this task: Jin et al. [12] use Faster R-CNN reporting a bounding box mean average precision (mAP) in terms of intersection over union (IoU) of 63.1%, while Kletz et al. [16] report a similar mAP of 64.5% for bounding boxes and in addition, since they are using the segmentation-capable Mask R-CNN, a segmentation mAP of 54.3%. Furthermore, in addition to surgical instruments, Zadeh et al. [43] attempt to localize anatomical regions, also using Mask R-CNN and achieving high segmentation mAPs for uterus (84.5%), ovary (29.6%) and tools (54.5%, similar to Kletz et al. [16]). Finally, Gibson et al. [5] define a deep residual network for supervised liver segmentation yielding a median Dice similarity score of 95%. Fu et al. [4] even show that this result score can be improved to 96.4% by following a different model training strategy, i.e. using a semi-supervised mean teacher method.

3 Methodology

Our approach targeting endometriosis detection and segmentation is summarized in Fig. 2. We start by gradually developing ENID by revising publicly released dataset GLENDa [20] and creating intermediate dataset GLENDa-VIS in Section 3.1. Subsequently, we thoroughly describe all employed data augmentation techniques such as

Fig. 2 Methodology overview – dataset creation, augmentations and deep model training

blurring, cropping and desaturation but also reflection removal as well as tracking in Section 3.2. Finally, after outlining our data partitioning approaches, Section 3.3 details our model training strategies and Section 3.4 lists our evaluation results.

3.1 Datasets

When considering all potential endometriosis locations and severities using both, the rASRM as well as the Enzian system, a complete dataset would have to include well over 50 classes, which would be challenging to collect in large enough quantities. Therefore, the publicly available dataset GLEND A [20] follows the strategy of disregarding lesion size altogether and focusing on three common locations as well as one specific type of endometriosis. Figure 3 depicts example images of the dataset's classes: peritoneum, ovary, uterus and deep infiltrating endometriosis³ (DIE). As indicated with a light green color, in every image at least one region is annotated using rectangular bounding boxes or polygon shapes. All of these annotations fully enclose the lesions, albeit mostly do not mark their exact boundaries. This is due to the difficulty of reasonably identifying such boundaries: often, pathologic areas contain multiple very small lesions that are infeasible to annotate individually (c.f. Fig. 3d, second image) or the borders of affected tissue blend in with adjacent healthy tissue (c.f. Fig. 3a, second image). Apart from this annotation strategy likely already entailing consequences for model training, an arguably much larger impact factor can be identified when regarding the visual dissimilarity of annotated regions within classes. Although partially showing similar patterns across classes as well as within them, the lesions nevertheless can also vary strongly, which is a highly detrimental characteristic for convolutional deep networks attempting to learn visually distinct features across classes.

³ Deep infiltrating endometriosis is found several millimeters into the lining of the abdominal cavity, i.e. the peritoneal tissue.

Fig. 3 GLENDa dataset class overview

Addressing these problems, we perform two steps for improvement. First, we reorganize the dataset's classes by disregarding lesion location and grouping pathologic areas by their visual similarity. This is accomplished by using most of GLENDa's images and altering their annotations accordingly. Examples of the resulting visual GLENDa dataset (GLENDa-VIS) are illustrated in Fig. 4. GLENDa-VIS contains four classes:

Mucus white-toned mucous areas, often part of adhesions or sclerotic areas

Vesicles clear blister-like areas, often resembling water drops

Implants dark-toned growths, often are surrounded by white sclerotic areas

Abnormal Tissue irritated/bleeding tissue, e.g. increased/deformed blood vessels or bleeding endometrial cells

GLENDa-VIS is a preliminary dataset with annotations created exclusively from rectangles that can overlap to form simple polygonal areas. It is intentionally created in this rudimentary manner, as for the purposes of this work it is merely used to determine the most promising visual endometriosis characteristic for the second improvement step performed: creating a larger binary dataset using the best performing GLENDa-VIS class.

After evaluating the instance segmentation performance of both GLENDa and GLENDa-VIS (c.f. Tables 9 and 10), we identify Implants as the most promising candidate for our final Endometrial Implants Dataset (ENID), which is shown exemplarily in Fig. 5. We base most of our evaluations in this work on ENID, in which we finally

Fig. 4 GLEND-VIS dataset class overview

Fig. 5 ENID dataset examples

address the above issue of annotation imprecision – lesions are enclosed by carefully freehand-drawn polygons created using the Endoscopic Concept Annotation Tool [25] and according to following criteria:

color Implants generally appear as dark, sometimes dotted areas with varying color tones. Annotators predominantly focus on clearly visible implants.

size Annotated implants are required to be of a certain, well visible size – very small lesions are very likely not have a significant impact.

boundary Annotations are created precisely, carefully identifying lesion boundaries. Here annotators focus on the transition from darker to lighter tissue.

Table 1 Dataset Statistics Comparison – number of patient cases, images and annotations

Dataset		Cases	Images	Annotations
GLEND A	peritoneum	84	257	489
	ovary	27	53	54
	uterus	11	17	25
	die	29	55	59
	total	102	373	628
GLEND A-VIS	mucus	77	246	442
	vesicles	59	153	250
	implants	86	281	652
	abnormal tissue	41	66	85
	total	96	342	1485
ENID	implants	108	160	358

Similar to GLEND A/GLEND A-VIS, ENID is collected from a selection of frames belonging to a laparoscopic video archive of over 500 individual surgery recordings. Table 1 presents a comparison of statistics for all three datasets: number of patient cases, images and annotations.⁴ As indicated by the cases column, all datasets contain frames taken from close to 100 surgeries, which on one hand ensures data variety for increased generalizability and on the other hand lowers the risk of overfitting due to potential similarities in training and validation data. This specifically is deemed important for creating the single-class ENID dataset, which is collected from altogether 108 surgeries – more than the number of cases for for the combination of all classes in GLEND A or GLEND A-VIS.

3.2 Augmentation

Apart from comparing the instance segmentation performance of the raw datasets, we also investigate the performance impact of applying several augmentations to ENID. We altogether explore seven augmentation techniques, five of which are commonly used in the field of image retrieval. All augmentations are exclusively applied to the training portion of the dataset only, while the validation and test sets are left unchanged.

Figure 6 shows six applied augmentations, where Fig. 6a-e are applied in the following manner:

- For every image i of training split S_t , one or multiple augmentations are applied n times increasing S_t by a factor of $n + 1$, i.e. $S_t^a = S_t \cup \{i_1^a, \dots, i_n^a\}$. In all our experiments we use $n = 2$ effectively tripling the training datasets.
- Augmentations are applied in random intensities on a set interval of potential values differing for every augmentation type.

⁴ Please note that individual frames are allowed to contain annotations for multiple classes, hence, the cases/images totals for GLEND A/GLEND A-VIS in fact are lower than they would be when simply summing up all class statistics.

Fig. 6 Summary of applied augmentations – original (left), augmented (right)

In the following we describe corresponding image augmentations⁵ and their intensity scales:

rotating rotate images on a scale between $[-45, 45]$ degrees

blurring apply motion blur using random angles and directions with kernel sizes between $[3, 11]$ pixels

cropping crop each image side by a percentage between $[0.0, 0.25]$, input size is retained

perspective transform transform a frame's perspective by placing 4 points relative to its corners in a distance (scale) between $[0.01, 0.15]$, input size is retained

desaturation reduce image saturation by a percentage between $[0.15, 0.40]$

The remaining augmentation shown in Fig. 6f represents a more special case, hence, it is applied differently. The introduction of bright light sources for illumination during laparoscopic surgery frequently creates specular reflections, especially on wet surfaces or bulging tissue as is formed in the case of shallow endometrial implants. This can have detrimental effects on the learning performance (cf. Fig. 8d), since it may lead to associating reflections with lesions. Therefore, we explore the effects of automatically removing very bright reflections, i.e. regions above a certain brightness threshold, and correct the images using inpainting as proposed by Telea et al. [36]. As we want to remove reflections altogether from S_t , when applying this augmentation we simply replace these training images with the processed ones, i.e. altering instead of enlarging the corresponding data. Although removing potentially important information using such an approach, we, nevertheless, expect it to mitigate the negative performance impact of specular reflections.

⁵ For more detailed information about these standard augmentations please refer to the utilized library `imgaug` [13].

Fig. 7 Tracked ENID examples, one tracked annotation per row (GT marks original ground truth annotation)

For our final augmentation, we perform object tracking using the annotated frames together with their source videos. We accomplish this task by using kernelized correlation filters [10], a simple and fast bounding box tracking methodology. We process each annotation on a frame-to-frame basis, i.e. every successfully tracked region is used for tracking the next region. Albeit simple bounding boxes can not capture the exact ground truth region or its deformation, we nevertheless achieve good results by simply relocating and resizing the original annotations according to the tracked bounding boxes – Fig. 7 shows some of the results. Subsequently, we manually review all of the tracked frames and remove evidently incorrect results. Following this approach, we are able to augment the full dataset approximately by a factor of 81 (160 vs. 12981 frames). The final augmentation factor, however, depends on how the dataset is split into training, validation and test set, since similar to all other augmentations this one is as well only applied to the training portion. Also, in order to avoid feeding the network with almost identical images, as a last step we uniformly sample tracked frames with three different frame intervals: $v=10$, $v=20$ and $v=30$. For each interval, starting with the ground truth frame, we pick every v -th frame from a reviewed tracking sequence and repeating this process for the entire training set, yielding three augmented datasets: tracked_{10} , tracked_{20} and tracked_{30} .

Finally, in order to compromise between swift model training and result expressiveness, we use two dataset splits throughout all our experiments (training/validation/test): 60/20/20 and 80/10/10. For splitting, special attention is paid to avoid similar frames across splits and risk overfitting: patient cases, i.e. full surgery recordings, can only be part of a single split portion. For brevity, we only include result listings for the 60/20/20 split throughout the paper, while listing everything related to 80/10/10 in the [Appendix](#), as to provide the interested reader with further insights. Accordingly, Table 2 lists augmentation statistics related to the 60/20/20 split and Table 8 represents its [Appendix](#)-counterpart for 80/10/10. When regarding the number of images in the augmented training splits we can conclude that, while tripling our data with basic augmentations, we achieve an augmentation factor of over 7 with tracking. Additionally, convincing findings in our evaluations (cf. Section 3.4) also encourage us to combine augmentations, which are listed as the last four entries in the TRAINING section of Table 2. First, we investigate the effects of combining all simple augmentations for non-tracked (all w/o tracking) as well as tracked (all w/ tracked_{30}) data and second, discuss other promising augmentation combinations per split

Table 2 ENID (60/20/20) – augmentation statistics, cases, images and annotations

Augmentation	Cases	Images	Annotations
TRAINING			
None	64	96	240
Rotating	64	288	713
Blurring, desaturation	64	288	720
Cropping	64	288	672
Perspective transform	64	288	703
Reflection removal	64	96	240
Tracked ₁₀	64	653	1110
Tracked ₂₀	64	361	649
Tracked ₃₀	64	265	486
All w/o tracking	64	288	637
All w/ tracked ₃₀	64	786	1292
Blurring & rotating	64	288	713
Cropping & rotating	64	288	658
VALIDATION			
All	24	32	61
TESTING			
All	20	32	57

(blurring/cropping/rotating). Analogously to the above descriptions, we keep using $n = 2$ for these additional approaches and apply all involved augmentations in a sequence to both images. Lastly, it is worth mentioning that, albeit the number of images stays consistent across augmentations, the number of annotations varies because they potentially are removed by geometrical image transformations. Additionally, if they are located near an image border a transformation can as well possibly dissect them. As a consequence during the augmentation process, we verify that every augmented image at least contains one valid annotation.

3.3 Model training

Having developed the datasets and prepared ENID in the above described way, we at this point are ready for model training. As mentioned in the introductory section, we use both Faster R-CNN (Faster) as well as Mask R-CNN (Mask) as architectures, each trained with ResNet-50 (R-50) and ResNet-101 (R-101) as backbones. We conduct transfer learning, i.e. the models are initialized with pre-trained weights from the COCO dataset [21], while adjusting the last layer for our detection and segmentation purposes: since we want to compare the performance of single lesion classes, we exclusively train our models to give binary output. To achieve this, we correspondingly split up all multi-class datasets GLEND A and GLEND A-VIS into binary datasets of four classes each. The utilized loss functions incorporate class (log loss), bounding box (smooth L_1 loss) and for Mask R-CNN pixel-based segmentation (binary cross entropy loss) predictions as described in [6, 9]. Additionally, for optimization we choose stochastic gradient descent (SGD) and conduct

initial experiments for finding the optimal learning rate $\eta \in \{0.01, 0.001, 0.0001\}$ with the result of choosing $\eta = 0.001$ due to its overall superior performance. We also conduct preliminary experiments using Adam, yet, find that this optimization strategy tends to perform better for classifying rather than segmenting endometriosis. Training is conducted for a fixed number of 50 epochs (approx. 25.5K iterations) and we create a model checkpoint approximately every 5 epochs.

The utilized networks are implemented in Python using Detectron2⁶ – an object detection framework powered by the PyTorch⁷ deep learning framework. Regarding hardware, a workstation running Linux Ubuntu 18.x with the following specifications was used: Intel Core i7-5820K CPU @ 3.30GHz x 6, 32 GiB DDR3 @ 1333 MHz, Nvidia GeForce GTX 1080. Depending on the amount of input data, training a single network for 50 epochs requires approximately 2–4h to complete.

3.4 Evaluation

This section contains all the results achieved for evaluating our trained models (Section 3.4.1) as well as qualitative prediction insights (Section 3.4.2). Similarly to Section 3.2, we only report results for the 60/20/20 split, while listing the ones for 80/10/10 in the Appendix. Additionally, in order to further condense table sizes, for every potentially augmented dataset, we merely report the best performing network combination (R-CNN + backbone) evaluated at the epoch checkpoint that achieves the highest mean average precision (mAP) value.

3.4.1 Quantitative results

Using the COCO-detection [21] metrics, we evaluate a number of mAP values for different intersection over union (IoU) thresholds – the higher the threshold, the more a predicted area is required to overlap with a given ground truth to be considered correct. Since we are using Faster R-CNN and Mask R-CNN architectures, we report bounding box⁸ mAPs (mAP^{bb}) as well as pixel mask segmentation mAPs (mAP). IoU thresholds are identified by subscript numbers: mAP_{50-95} indicates the mAP calculated over 10 thresholds starting from 50% percent up to 95% IoU overlap using intervals of 5%, while mAP_{50} and mAP_{75} describe the mAP at single specific thresholds 50% as well as 75%.

First, we evaluate the performance of the raw (unaugmented) datasets GLENDa, GLENDa-VIS and ENID listing mAP comparisons for bounding box prediction in Table 3. Throughout GLENDa as well as GLENDa-VIS, we discover very low precisions of at most close or equal to 3% ($\text{mAP}_{50-95}^{\text{bb}}=0.030$). For many classes (ovary, uterus, die, abnormal tissue), result precisions for any trained models even are almost 0% ($\text{mAP}_{50-95}^{\text{bb}} < 0.000$). This is as well reflected in the 80/10/10 split (cf. Table 9) with the exception of ovary scoring around 10% ($\text{mAP}_{50-95}^{\text{bb}}=0.107$). When specifically observing GLENDa-VIS, we identify vesicles ($\text{mAP}_{50-95}^{\text{bb}}=0.030$) and implants ($\text{mAP}_{50-95}^{\text{bb}}=0.029$) as the best performing classes and their similar performance is confirmed by regarding the evaluations for

⁶ <https://github.com/facebookresearch/detectron2>

⁷ <https://pytorch.org/>

⁸ Please note that for Mask vs. Faster R-CNN bounding boxes we essentially are comparing RoI Pooling vs. RoI Align, the latter of which is not suffering from quantization loss, hence, is expected to yield more precise results.

Table 3 GLENDa, GLENDa-VIS, ENID (60/20/20) – bounding box prediction precision, raw; best (**bold**) and worst (*italic*)

Dataset		Arch.	Backb.	Ep.	mAP_{50-95}^{bb}	mAP_{50}^{bb}	mAP_{75}^{bb}
GLENDa	peritoneum	Mask	R-101	24	0.030	0.068	0.010
	ovary	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
	uterus	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
	die	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
GLENDa-VIS	mucus	Mask	R-101	35	0.015	0.026	0.020
	vesicles	Mask	R-101	30	0.030	0.068	0.004
	implants	Mask	R-101	35	0.029	0.069	0.020
	abn. tissue	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
ENID	implants	Mask	R-50	34	0.308	0.561	0.276

the 80/10/10 split (cf. Table 9). ENID being the more refined endometrial implants dataset outperforms both other datasets significantly with a precision over 30% ($mAP_{50-95}^{bb}=0.308$), which is even increased to over 50% ($mAP_{50}^{bb}=0.561$) when merely considering a 50% IoU overlap threshold. When regarding the mask segmentation results listed in Table 4, we observe similar results: GLENDa as well as GLENDa-VIS classes yield equally low or even lower segmentation scores, while ENID on the other hand scores significantly higher with over 30% for all thresholds ($mAP_{50-95}=0.309$) or above 50% for a 50% IoU threshold ($mAP_{50-95}=0.581$). Albeit performing slightly worse ($mAP_{50-95}^{bb}=0.288$, $mAP_{50}^{bb}=0.5$, $mAP_{50-95}=0.250$, $mAP_{50}=0.522$), the same trend can be observed for split 80/10/10 (cf. Tables 9 and 10).

Having observed superior results for ENID in all mAP categories, we finally compare the raw dataset's performance against the effects of applying different augmentations. Table 5 lists all precision values for bounding box predictions and only few of them seem to have a slight impact on the results. While cropping, perspective transform and desaturation yield very similar results to the raw dataset (0–0.5% improvement in mAP_{50-95}^{bb}), we discover that the best performing augmentations for bounding box prediction are blurring and rotating with 0.7–2.5% improvements in mAP_{50-95}^{bb} . Albeit these improvements can not be regarded as significant, in particular since they can not be observed for the 80/10/10 split

Table 4 GLENDa, GLENDa-VIS, ENID (60/20/20) – mask segmentation prediction precision, raw; best (**bold**) and worst (*italic*)

Dataset		Arch.	Backb.	Ep.	mAP_{50-95}	mAP_{50}	mAP_{75}
GLENDa	peritoneum	Mask	R-101	24	0.025	0.074	0.004
	ovary	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
	uterus	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
	die	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
GLENDa-VIS	mucus	Mask	R-101	35	0.015	0.026	0.020
	vesicles	Mask	R-101	30	0.029	0.068	0.027
	implants	Mask	R-101	35	0.028	0.069	0.019
	abn. tissue	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
ENID	implants	Mask	R-50	39	0.309	0.581	0.325

Table 5 ENID (60/20/20) – bounding box prediction precision, raw vs. augmented; best (**bold**) and worst (*italic*)

Augmentation	Arch.	Backb.	Ep.	mAP _{50–95} ^{bb}	mAP ₅₀ ^{bb}	mAP ₇₅ ^{bb}
None	Mask	R-50	34	0.308	0.561	0.276
Rotating	Mask	R-101	24	0.333	0.633	0.322
Blurring	Mask	R-50	44	0.315	0.582	0.309
Cropping	Mask	R-50	44	0.308	0.544	0.269
Perspective transform	Mask	R-101	44	0.311	0.591	0.292
Desaturation	Mask	R-50	49	0.313	0.566	0.303
Reflection removal	Mask	R-101	44	0.281	0.544	0.274
Tracking ₁₀	Faster	R-101	15	<i>0.234</i>	<i>0.460</i>	<i>0.183</i>
Tracking ₂₀	Mask	R-50	40	0.268	0.516	<i>0.183</i>
Tracking ₃₀	Mask	R-50	50	0.270	0.515	0.232
All w/o tracking	Mask	R-101	19	0.299	0.590	0.231
all w/ tracked ₃₀	Faster	R-101	19	0.276	0.526	0.216
Blurring & rotating	Mask	R-101	24	0.307	0.638	0.250
Cropping & rotating	Mask	R-101	24	0.344	0.627	0.278

– here all augmentations even perform worse than the raw dataset (cf. Table 11). Looking at the mask segmentation results, however, we discover a slight impact of rotating for both, split 60/20/20 (0.320 vs. 0.309, cf. Table 6) as well as split 80/10/10 (0.253 vs. 0.250, cf. Table 12). The latter split even indicates a performance boost of 2.3–3.85% for cropping in all listed mAP values compared to the raw dataset performance. Further, for the combination of all augmentations applied to the raw and tracking₃₀ data, we observe similar performance to most individually augmented data – for 80/10/10 we even discover many of the worst-performing mAPs for both bounding box and mask predictions. Lastly, it is worth noting that although the tracking approach consistently decreases the performance,

Table 6 ENID (60/20/20) – mask segmentation prediction precision, raw vs. augmented; best (**bold**) and worst (*italic*)

Augmentation	Arch.	Backb.	Ep.	mAP _{50–95}	mAP ₅₀	mAP ₇₅
None	Mask	R-50	39	0.309	0.581	0.325
Rotating	Mask	R-50	24	0.320	0.640	0.255
Blurring	Mask	R-50	44	0.290	0.582	0.289
Cropping	Mask	R-101	24	0.298	0.580	0.250
Perspective transform	Mask	R-101	14	0.283	0.573	0.267
Desaturation	Mask	R-50	19	0.286	0.575	0.229
Reflection removal	Mask	R-101	24	0.286	0.547	0.244
Tracking ₁₀	Mask	R-101	30	<i>0.197</i>	0.489	<i>0.102</i>
Tracking ₂₀	Mask	R-101	10	0.228	<i>0.465</i>	0.175
Tracking ₃₀	Mask	R-101	15	0.217	0.473	0.119
All w/o tracking	Mask	R-101	19	0.277	0.572	0.229
All w/ tracked ₃₀	Faster	R-101	14	0.250	0.487	0.214
Blurring & rotating	Mask	R-101	19	0.294	0.608	0.265
Cropping & rotating	Mask	R-101	29	0.324	0.642	0.255

the results indicate that more samples generally perform worse, i.e. tracked_{20} and tracked_{30} mostly show higher mAP values than tracked_{10} with the exception of $\text{mAP}_{50-95}^{\text{bb}}$ for split 80/10/10 (cf. Tables 5 and 11)

Based on above observations, we identify three augmentations that generally seem to have an impact on the resulting precision values, whether for bounding box prediction or mask segmentation on either of the splits: rotating, cropping and blurring. Therefore, we choose to combine these particular augmentations in order to determine, if the outcomes can further be improved. After training additional models on ENID that is augmented by jointly applying above techniques, we subsequently decide to add the following two augmentation combinations to our evaluations: blurring & rotating as well as cropping & rotating. Unfortunately, for both splits, the best models for augmenting ENID through blurring- and rotating are not improving prediction precision and, in fact, even show slightly detrimental effects on mask segmentation results for split 60/20/20 (0.294 vs. 0.309 mAP_{50-95} , cf. Table 6) and split 80/10/10 (0.238 vs. 0.250 mAP_{50-95} , cf. Table 12). On the other hand, models for cropping- & rotating-augmented ENID mostly outperform raw ENID models for bounding box prediction and segmentation in both splits, albeit merely by up to 3.6% precision. Nevertheless, this ultimately seems to be the most robust performance improvement indicator. Thus, we finally summarize the overall best augmentation techniques for any of the splits to be cropping, rotating and their combination. Subsequently, we put a stronger focus on the segmentation results of the 60/20/20 split and compare the best models trained on raw ENID to those trained on rotating- as well as cropping- & rotating-augmented ENID: although the performance improvement of applying augmentation in this case merely amounts to 1.1-1.5% mAP_{50-95} , the models' differences, nevertheless, become apparent when qualitatively inspecting their prediction outputs.

3.4.2 Qualitative results

A qualitative inspection of results taken from the 60/20/20 testing set reveals advantages as well as disadvantages in our evaluation's best performing models, which are chosen according to segmentation precision (mAP_{50-95}). Figure 8 compares selected prediction results of the best model for raw ENID ($M_{\text{raw}}^{\text{M-R-50}}$) with the best models trained on ENID with augmentations, i.e. rotation ($M_{\text{rot}}^{\text{M-R-101}}$) and cropping combined with rotation ($M_{\text{crop+rot}}^{\text{M-R-101}}$). Every row in the figure compares a test split ground truth annotation with predictions of all chosen models in the same order as mentioned above and using a confidence threshold of 50%. In contrast to the green outlined ground truth annotations, the model predictions are randomly colored and include pixel segmentation masks as well as bounding boxes for increased visibility.

Starting with Fig. 8a, $M_{\text{rot}}^{\text{M-R-101}}$ fails to predict some lesion areas, yet, according to the ground truth, correctly is not mispredicting the small dark spot on the right side of the image, in contrast to the other models. This region, however, is not as easily distinguishable from a pathologic area: in fact, ground truth annotations are only made for obvious lesions, therefore, the prediction in question could be considered suspicious and in need of further examination by an expert. Thus, arguably in this case, an expert might even prefer such suspicious areas to be included in the results, despite their potential of being false alarms. Subsequently, in all of the remaining figures, we examine the most common mispredictions discovered throughout all trained models. Figure 8b demonstrates how $M_{\text{raw}}^{\text{M-R-50}}$ and $M_{\text{rot}}^{\text{M-R-101}}$ confuse a darker background area for a lesion, while $M_{\text{crop+rot}}^{\text{M-R-101}}$ manages to avoid this mistake. In fact, this is discovered most frequently with differing severity, i.e.

Fig. 8 Selected qualitative result comparison of best performing models using a 50% confidence threshold – ground truth (left), Mask R-CNN with ResNet-50 on raw ENID ($M_{\text{raw}}^{\text{M-R-50}}$, middle left), Mask R-CNN with ResNet-101 on rotating-augmented ENID ($M_{\text{rot.}}^{\text{M-R-101}}$, middle right), Mask R-CNN with ResNet-101 on cropping- & rotating-augmented ENID ($M_{\text{crop+rot.}}^{\text{M-R-101}}$, right)

multi-shaded background areas potentially are mispredicted partially up to fully. Moreover, a further cause of mispredictions are prominent color transitions, as can be discovered in Fig. 8c, where $M_{\text{rot.}}^{\text{M-R-101}}$ confuses a blood vessel for a lesion. Blood or blood vessels are contrasting in color with their surrounding tissue, which makes them easy targets to mispredict, especially when embedded in mucus-like surrounding tissue as is the case in the figure. Another less obvious but prevalent misprediction type are reflections as shown in Fig. 8d, which demonstrates $M_{\text{raw}}^{\text{M-R-50}}$ falsely predicting a large specular reflection to be a lesion, while the other models behave correctly in this situation. This presumably happens because reflections are part of many implant annotations (cf. Fig. 5), hence, they are learned to be part of the lesion regions. Although mitigating this particular problem, reflection removal evidently increases the amount of other mispredictions, thus, altogether having a detrimental rather than beneficial effect (cf. Table 6). Surprisingly, at least in the depicted case, cropping seems to as well have a positive effect on this problem. Lastly,

instruments also exhibit attributes that can lead to confusion: many possess black shafts that under certain lighting conditions are easily confused for implants and their often silvery top parts frequently reflect light, which results in reflection mispredictions. For instance, Fig. 8e shows $M_{\text{raw}}^{\text{M-R-50}}$ struggling with instrument shafts, while the other models appear to be less susceptible to this kind of misprediction.

Overall, judging by the examples shown in Fig. 8, it seems apparent to deem $M_{\text{rot}}^{\text{M-R-101}}$ and $M_{\text{crop+rot}}^{\text{M-R-101}}$ as more robust than $M_{\text{raw}}^{\text{M-R-50}}$. However, it is worth again referring to their mere slight overall performance improvement of at most 1.5% mAP_{50-95} . Therefore, they ultimately many times are as well suffering from above mispredictions and lesion localization failures, yet, on different test images. Nevertheless, every above model seems to have its merits in certain situations, which may allow for increasing performance by applying them in sequence and taking decisions using majority voting.

4 Discussion

Reviewing the entirety of results, we observe several expected, yet, also unexpected outcomes. First, looking at the prediction performance of GLENDa, GLENDa-VIS and ENID we can confirm our conjecture that categorizing endometriosis by visual appearance rather than location is more suitable for state-of-the-art R-CNN networks. This particularly seems reasonable when focusing on a lesion manifestation like endometrial implants, which potentially can occur at any bodily region susceptible to endometriosis. Moreover, we ultimately base choosing to further investigate this particular class on observing its overall best performance as part of GLENDa-VIS, while taking into account two different dataset splits. This decision can arguably even stronger be justified by observing the somewhat exceptional performance of models trained on GLENDa's ovary class, which appears to be influenced by its inclusion of many implants-like annotations (cf. Figs. 3b or 7, first row). Nevertheless, ovary-trained models in direct comparison to ENID-trained models show significantly lower performance, which is attributable to the difference in annotation strategies used when creating the classes: apart from the necessity of consulting medical experts, precisely enclosing lesion regions as well as setting clear annotation guidelines are key components for largely increasing segmentation performance.

On the downside, we discover a strong decrease in precision when requiring predictions to have a higher lesion area overlap. While this is expected considering the difficulty of the problem, we on average identify a considerable decrease of close to 40% between results calculated with a 50% and 75% overlap requirement, which holds true for models trained on raw as well as augmented ENID. Surprisingly, most proposed augmentations do not yield any significant performance improvements, which, although standing in contrast to results achieved in traditional image segmentation tasks [23], we find to be in line with the discoveries in other medical fields, such as the findings of Fox et al., who apply Mask R-CNN for tool segmentation in cataract surgery using microscopy videos [3].

Nevertheless, when striving to identify the best performing augmentations by regarding quantitative as well as qualitative evaluations we identify simple rotation, the results of which can even further be improved in combination with cropping. Interestingly, this combination for various examples counteracts some of the major causes of mispredictions: dark background areas, blood vessels, specular reflections and surgical

instruments. Nevertheless, the best performing augmented models still merely show slight performance improvements over the non-augmented models. Interestingly, some mispredictions could be regarded as questionable and in need of further expert inspection, since it is possible that pathological regions are omitted during ground truth annotation, either because of not conforming to the annotation policy or being too inconclusive. Since larger numbers of such unannotated lesions could influence training, this particularly could explain the comparably poor performance of models trained on tracked data: single annotated frames do not necessarily contain all lesion regions of a whole scene, however, during tracking they can easily be brought into view by camera movements or surgery actions like moving an organ to reveal occluded implants. Finally, it is also possible that, despite uniform sampling, tracked images still are too similar, causing a disproportionate amount of weight to be put on easily trackable lesion scenes while decreasing the impact of others.

5 Contributions

As outlined in Section 2 at the time of writing, we are unable to retrieve any directly comparable work on endometriosis segmentation. Nevertheless, we list some loosely related work in Table 7, where we set our work in contrast to classification, object detection and segmentation tasks on endoscopy [12, 16, 40, 43] as well as microscopy [3] datasets.

For classification, we list Visalaxi et al. [40] who achieve a very high binary endometriosis classification accuracy of 90.0% using ResNet on the GLENDa [20] dataset. When compared to our work, besides classification being a different task, with defining endometrial implants we additionally are targeting a more confined type of lesion than the various ones included in GLENDa. Additionally, it remains unclear how the dataset was used or

Table 7 Comparison of loosely related work targeting Instruments (Instr.), Anatomy (Anat.) and Endometriosis (Endom.). Results are reported for classification (accuracy), object detection (mAP_{50-95}^{bb} , mAP_{50}^{bb}) and segmentation (mAP_{50-95} , mAP_{50})

Reference	Topic	Data	Method	Results	
CLASSIFICATION				accuracy	
Visalaxi et al. [40]	Endom.	[20]	ResNet	90.0	
OBJECT DETECTION				mAP_{50-95}^{bb}	mAP_{50}^{bb}
Jin et al. [12]	Instr.	[12]	Faster R-CNN	–	63.1
Fox et al. [3]	Instr.	[7]	Mask R-CNN	59.2	83.9
Kletz et al. [16]	Instr.	custom	Mask R-CNN	64.5	83.1
Ours (cropping & rotating)	Endom.	[20]/ENID	Mask R-CNN	34.4	62.7
SEGMENTATION				mAP_{50-95}	mAP_{50}
Zadeh et al. [43]	Anat.	custom	Mask R-CNN	–	84.5/29.6
Zadeh et al. [43]	Instr.	custom	Mask R-CNN	–	54.5
Kletz et al. [16]	Instr.	custom	Mask R-CNN	54.3	81.4
Fox et al. [3]	Instr.	[7]	Mask R-CNN	40.9	71.0
Ours (cropping & rotating)	Endom.	[20]/ENID	Mask R-CNN	32.4	64.2

processed. Specifically, the authors used 6000/25682 images for their evaluations but omit how many of them show pathology as well as which lesion types are included.

For object detection, we contrast our work against the task of localizing instruments. Albeit we achieve comparable results to Jin et al. [12], we generally discover that other approaches [3, 16] outperform ours by approximately doubling $\text{mAP}_{50-95}^{\text{bb}}$. The fact that both of these studies also used Mask R-CNN confirms our assumptions that it is much easier to detect and localize instruments than endometrial implants. This becomes even more apparent when visually comparing the prominent appearance of surgical instruments against a much more subtle endometriosis region, e.g. by observing Fig. 8e.

For segmentation, we similarly include studies targeting instruments, yet, also another study that additionally targets anatomy. Here we discover a lower performance gap to both above-addressed studies and even reach almost comparable performance to Fox et al. [3]. However, since this work is using microscopy images, we argue that it is the overall least comparable study besides Visalaxi et al. [40]. Although Zadeh et al. [43] are missing mAP_{50-95} values, they achieve good performance for uterus segmentation (84.5% mAP_{50}), yet, much lower results for uterus (29.6% mAP_{50}) and instrument segmentation (54.0% mAP_{50}). This is surprising, given that instruments should be more distinguishable from any anatomical structure. Nevertheless, it underlines our main conception that segmentation strongly depends on the distinct visual appearance of a given target object.

Finally, given the overall absence of directly comparable work, we believe that this study makes an important step in the direction of automatic endometriosis segmentation in the domain of laparoscopic gynecology. Specifically, reorganizing medically accepted endometriosis classes according to their visual similarity shows a large performance improvement. Therefore, we consider this study as well as the accompanying publicly released endometrial implant dataset (ENID) to be of particular importance for advancing computer-aided endometriosis analysis.

6 Conclusion

We target a visually-oriented approach for endometriosis segmentation in laparoscopic surgery videos utilizing a publicly available dataset to develop a novel endometrial implants dataset including precise hand-drawn annotations. Further, we evaluate this dataset using Faster as well as Mask R-CNN combined with different ResNet backbones. Results show a large performance increase compared to the public dataset, yet merely slight improvements when applying geometrical data augmentations. The best results are achieved using simple random image rotation combined with cropping, which surprisingly counteracts some of the major causes of mispredictions. Establishing a baseline for future work in this domain, we publish the novel dataset as well as selected deep models with this study. For ensuing research, we plan on further investigating data augmentations and their limited applicability to laparoscopic videos as well as adding additional classes to the currently binary dataset. Finally, we believe this study lays important groundwork for the topic of computer-aided endometriosis lesion detection and we hope it inspires other researchers to contribute to this yet scarcely researched topic.

Appendix

This Appendix contains all statistics and evaluations for ENID split according to percentages 80/10/10 (training/validation/test).

Table 8 ENID, 80/10/10 – augmentation statistics, cases, images and annotations

Augmentation	Cases	Images	Annotations
TRAINING			
None	83	128	279
Rotating	83	384	832
Blurring, desaturation	83	384	837
Cropping	83	384	809
Perspective transform	83	384	826
Reflection removal	83	128	279
Tracked ₁₀	83	1095	1654
Tracked ₂₀	83	595	937
Tracked ₃₀	83	419	673
Blurring & rotating	83	384	834
Cropping & rotating	83	384	788
All w/o tracking	64	384	751
All w/ tracked ₃₀	64	1245	1841
VALIDATION			
All	12	16	32
TESTING			
All	13	16	47

Table 9 GLENDa, GLENDa-VIS, ENID (80/10/10) – bounding box prediction precision, raw; best (**bold**) and worst (*italic*)

Dataset		Arch.	Backb.	Ep.	mAP _{50–95} ^{bb}	mAP ₅₀ ^{bb}	mAP ₇₅ ^{bb}
GLENDa	peritoneum	Faster	R-101	45	0.028	0.078	0.003
	ovary	Mask	R-101	45	0.107	0.119	0.119
	uterus	all	all	–	<i><0.000</i>	<i><0.000</i>	<i><0.000</i>
	die	Mask	R-50	50	0.036	0.119	<i><0.000</i>
GLENDa-VIS	mucus	Faster	R-101	35	0.029	0.053	0.04
	vesicles	Mask	R-101	39	0.046	0.105	0.03
	implants	Faster	R-50	45	0.054	0.127	0.028
	abn. tissue	Mask	R-101	49	0.008	0.016	<i><0.000</i>
ENID	implants	Mask	R-101	39	0.288	0.5	0.332

Table 10 GLENDa, GLENDa-VIS, ENID (80/10/10) – mask segmentation prediction precision, raw; best (**bold**) and worst (*italic*)

Dataset		Arch.	Backb.	Ep.	mAP _{50–95}	mAP ₅₀	mAP ₇₅
GLENDa	peritoneum	Mask	R-101	15	0.015	0.037	<0.000
	ovary	Mask	R-101	45	0.048	0.119	<0.000
	uterus	Mask	all	-	<0.000	<0.000	<0.000
	die	Mask	R-101	50	0.036	0.119	<0.000
GLENDa-VIS	mucus	Mask	R-101	35	0.023	0.062	0.021
	vesicles	Mask	R-101	44	0.042	0.100	0.033
	implants	Mask	R-101	20	0.048	0.137	0.033
	abn. tissue	Mask	R-101	49	0.010	0.016	0.016
ENID	implants	Mask	R-50	39	0.250	0.522	0.195

Table 11 ENID (80/10/10) – bounding box prediction precision, raw vs. augmented; best (**bold**) and worst (*italic*)

Augmentation	Arch.	Backb.	Ep.	mAP _{50–95} ^{bb}	mAP ₅₀ ^{bb}	mAP ₇₅ ^{bb}
None	Mask	R-101	39	0.288	0.500	0.332
Rotating	Mask	R-101	44	0.287	0.440	0.395
Blurring	Mask	R-101	24	0.249	0.459	0.271
Cropping	Mask	R-50	44	0.266	0.499	0.216
Perspective transform	Faster	R-101	24	0.249	0.497	0.207
Desaturation	Mask	R-101	29	0.252	0.397	0.306
Reflection removal	Mask	R-101	49	0.253	0.417	0.317
Tracking ₁₀	Mask	R-101	20	0.269	0.432	0.321
Tracking ₂₀	Mask	R-101	10	0.240	0.459	0.206
Tracking ₃₀	Mask	R-50	35	0.248	0.497	0.227
All w/o tracking	Mask	R-101	44	<i>0.231</i>	0.388	0.234
All w/ tracked ₃₀	Mask	R-101	30	0.236	0.472	<i>0.159</i>
Blurring & rotating	Mask	R-50	29	0.257	0.479	0.260
Cropping & rotating	Mask	R-101	49	0.261	0.472	0.278

Table 12 ENID (80/10/10) – mask segmentation prediction precision, raw vs. augmented; best (**bold**) and worst (*italic*)

Augmentation	Arch.	Backb.	Ep.	mAP _{50–95}	mAP ₅₀	mAP ₇₅
None	Mask	R-50	39	0.250	0.522	0.195
Rotating	Mask	R-101	44	0.253	0.440	0.231
Blurring	Mask	R-101	24	0.242	0.434	0.253
Cropping	Mask	R-50	44	0.277	0.545	0.258
Perspective transform	Mask	R-50	49	0.252	0.472	0.183
Desaturation	Mask	R-50	34	0.241	0.492	0.196
Reflection removal	Mask	R-101	29	0.228	0.421	0.215
Tracking ₁₀	Mask	R-101	25	0.226	0.429	0.195
Tracking ₂₀	Mask	R-50	25	0.236	0.466	0.157
Tracking ₃₀	Mask	R-50	35	0.246	0.457	0.185
All w/o tracking	Mask	R-101	44	<i>0.200</i>	<i>0.374</i>	0.213
All w/ tracked ₃₀	Mask	R-101	30	0.217	0.472	<i>0.099</i>
Blurring & rotating	Mask	R-50	29	0.238	0.479	0.179
Cropping & rotating	Mask	R-101	49	0.253	0.472	0.228

Acknowledgements This work was funded by the FWF Austrian Science Fund under grant P 32010-N38.

Funding Open access funding provided by University of Klagenfurt.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Canis M, Donnez J, Guzick D, Halme J, Rock J, Schenken R, Vernon M (1997) Revised american society for reproductive medicine classification of endometriosis: 1996. Fertility and Sterility 67(5):817–821. [https://doi.org/10.1016/S0015-0282\(97\)81391-X](https://doi.org/10.1016/S0015-0282(97)81391-X)
2. Du W, Rao N, Liu D, Jiang H, Luo C, Li Z, Gan T, Zeng B (2019) Review on the applications of deep learning in the analysis of gastrointestinal endoscopy images. IEEE Access 7:142053–142069
3. Fox M, Taschwer M, Schoeffmann K (2020) Pixel-based tool segmentation in cataract surgery videos with mask r-cnn. In: 2020 IEEE 33rd international symposium on computer-based medical systems (CBMS), IEEE, pp 565–568
4. Fu Y, Robu MR, Koo B, Schneider C, van Laarhoven S, Stoyanov D, Davidson B, Clarkson MJ, Hu Y (2019) More unlabelled data or label more data? a study on semi-supervised laparoscopic image segmentation. In: Domain adaptation and representation transfer and medical image learning with less labels and imperfect data, Springer, pp 173–180
5. Gibson E, Robu MR, Thompson S, Edwards PE, Schneider C, Gurusamy K, Davidson B, Hawkes DJ, Barratt DC, Clarkson MJ (2017) Deep residual networks for automatic segmentation of laparoscopic videos of the liver. In: Medical imaging 2017: Image-guided procedures, robotic interventions, and modeling, international society for optics and photonics, vol 10135, pp 101351M

6. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
7. Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, Quellec G, Chow A, Nehme J, Luengo I, Stoyanov D (2019) Cadis: Cataract dataset for image segmentation. [arXiv:1906.11586](https://arxiv.org/abs/1906.11586)
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, pp 770–778. <https://doi.org/10.1109/CVPR.2016.90>
9. He K, Gkioxari G, Dollár P, Girshick RB (2020) Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell* 42(2):386–397. <https://doi.org/10.1109/TPAMI.2018.2844175>
10. Henriques JF, Caseiro R, Martins P, Batista J (2014) High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(3):583–596
11. Jha D, Ali S, Johansen HD, Johansen DD, Rittscher J, Riegler MA, Halvorsen P (2020) Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning. [arXiv:2011.07631](https://arxiv.org/abs/2011.07631)
12. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L (2018) Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE winter conference on applications of computer vision (WACV), IEEE, pp 691–699
13. Jung AB, Wada K, Crall J, Tanaka S, Graving J, Reinders C, Yadav S, Banerjee J, Vecsei G, Kraft A, Rui Z, Borovec J, Vallentin C, Zhydenko S, Pfeiffer K, Cook B, Fernández I, De Rainville FM, Weng CH, Ayala-Acevedo A, Meudec R, Laporte M, et al. (2020) Imgaug. <https://github.com/aleju/imgaug>, Accessed 01 Feb 2020
14. Keckstein J, Hudelist G (2020) Classification of die including bowel endometriosis: from r-asmr to #enzian-classification. *Best Pract Res Clin Obstet Gynaecol*
15. Keckstein J, Ulrich U, Possover M, Schweppe K et al (2003) Enzian-klassifikation der tief infiltrierenden endometriose. *Zentralblatt für Gynäkologie* 125:291
16. Kletz S, Schoeffmann K, Benois-Pineau J, Husslein H (2019) Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In: 2019 International conference on content-based multimedia indexing (CBMI), IEEE, pp 1–6
17. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems* 25, Curran Associates, Inc., pp 1097–1105. <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
18. Leibetseder A, Primus MJ, Petscharnig S, Schoeffmann K (2017) Real-time image-based smoke detection in endoscopic videos. In: Wu W, Yang J, Tian Q, Zimmermann R (eds) *Proceedings of the on thematic workshops of ACM multimedia 2017*, October 23 – 27, 2017. ACM, Mountain View, CA, USA, pp 296–304. <https://doi.org/10.1145/3126686.3126690>
19. Leibetseder A, Petscharnig S, Primus MJ, Kletz S, Münzer B, Schoeffmann K, Keckstein J (2018) Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In: César P, Zink M, Murray N (eds) *Proceedings of the 9th ACM multimedia systems conference, MMSys 2018*, June 12–15, 2018. ACM, Amsterdam, The Netherlands, pp 357–362. <https://doi.org/10.1145/3204949.3208127>
20. Leibetseder A, Kletz S, Schoeffmann K, Keckstein S, Keckstein J (2020) GLEND: gynecologic laparoscopy endometriosis dataset. In: Ro YM, Cheng W, Kim J, Chu W, Cui P, Choi J, Hu M, Neve WD (eds) *MultiMedia Modeling - 26th International Conference, MMM 2020*, Daejeon, South Korea, January 5–8, 2020, *Proceedings, Part II*, Springer, *Lecture Notes in Computer Science*, vol 11962, pp 439–450. https://doi.org/10.1007/978-3-030-37734-2_36
21. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: *European conference on computer vision*, Springer, pp 740–755
22. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI (2017) A survey on deep learning in medical image analysis. *Medical Image Analysis* 42:60–88
23. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D (2020) Image segmentation using deep learning: A survey. [arXiv:2001.05566](https://arxiv.org/abs/2001.05566)
24. Münzer B, Schoeffmann K, Böszörményi L (2017) Content-based processing and analysis of endoscopic images and videos: A survey. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-016-4219-z>
25. Münzer B, Leibetseder A, Kletz S, Schoeffmann K (2019) ECAT - endoscopic concept annotation tool. In: Kompatsiaris I, Huet B, Mezaris V, Gurrin C, Cheng W, Vrochidis S (eds) *MultiMedia Modeling - 25th International Conference, MMM 2019*, Thessaloniki, Greece, January 8–11, 2019,

- Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol 11296, pp 571–576. https://doi.org/10.1007/978-3-030-05716-9_48,
26. Nwoye CI, Mutter D, Marescaux J, Padoy N (2019) Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. *International Journal of Computer Assisted Radiology and Surgery* 14(6):1059–1067
 27. Ozawa T, Ishihara S, Fujishiro M, Kumagai Y, Shichijo S, Tada T (2020) Automated endoscopic detection and classification of colorectal polyps using convolutional neural networks. *TherapeutiC Advances in Gastroenterology* 13:1756284820910659
 28. Park SY, Sargent D (2016) Colonoscopic polyp detection using convolutional neural networks. In: Tourassi GD, Armato SG (eds) *Medical Imaging 2016: Computer-Aided Diagnosis*, San Diego, California, United States, 27 February - 3 March 2016, International Society for Optics and Photonics, pp 978528. <https://doi.org/10.1117/12.2217148>
 29. Petscharnig S, Schöffmann K (2018) Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools and Applications* 77(7):8061–8079
 30. Piccialli F, Di Somma V, Giampaolo F, Cuomo S, Fortino G (2021) A survey on deep learning in medicine: Why, how and when? *Information Fusion* 66:111–137
 31. Rai HM, Chatterjee K, Gupta A, Dubey A (2020) A novel deep cnn model for classification of brain tumor from mr images. In: 2020 IEEE 1st international conference for convergence in engineering (ICCE), pp 134–138. <https://doi.org/10.1109/ICCE50343.2020.9290740>
 32. Rai HM, Chatterjee K, Dashkevich S (2021) Automatic and accurate abnormality detection from brain mr images using a novel hybrid unetresnext-50 deep cnn model. *Biomedical Signal Processing and Control* 66:102477
 33. Ren S, He K, Girshick R, Sun J (2017) Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
 34. Saraswat L, Ayansina D, Cooper K, Bhattacharya S, Horne A, Bhattacharya S (2018) Impact of endometriosis on risk of further gynaecological surgery and cancer: a national cohort study. *BJOG : An International Journal of Obstetrics & Gynaecology* 125(1):64–72
 35. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1–9
 36. Telea A (2004) An image inpainting technique based on the fast marching method. *Journal of Graphics Tools* 9(1):23–34
 37. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging* 36(1):86–97. <https://doi.org/10.1109/TMI.2016.2593957>
 38. Visalaxi S, Punnoose D, Muthu TS (2021a) An analogy of endometriosis recognition using machine learning techniques. In: 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV), IEEE, pp 739–746
 39. Visalaxi S, Punnoose D, Muthu TS (2021b) Lesion extraction of endometriotic images using open computer vision. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), IEEE, pp 747–751
 40. Visalaxia S, Muthua TS (2021) Automated prediction of endometriosis using deep learning. *Int J Nonlinear Anal Appl* 12(2):2403–2416
 41. Yang C, Zhao Z, Hu S (2020) Image-based laparoscopic tool detection and tracking using convolutional neural networks: a review of the literature. *Computer Assisted Surgery* 25(1):15–28
 42. Yengera G, Mutter D, Marescaux J, Padoy N (2018) Less is more: Surgical phase recognition with less annotations through self-supervised pre-training of cnn-lstm networks. [arXiv:180508569](https://arxiv.org/abs/180508569)
 43. Zadeh SM, Francois T, Calvet L, Chauvet P, Canis M, Bartoli A, Bourdel N (2020) Surgai: deep learning for computerized laparoscopic image understanding in gynaecology. *Surgical Endoscopy* 34(12):5377–5383