2022 26th International Conference on Pattern Recognition (ICPR)
August 21-25, 2022, Montréal, Québec, Canada

Pixel-accurate Segmentation of Surgical Tools based on Bounding Box Annotations

George Leifman
Google Research

Amit Aides
Google Research

Tomer Golany
Google Research

Daniel Freedman
Google Research

Ehud Rivlin
Google Research

2022 26th International Conference on Pattern Recognition (ICPR) | 978-1-6654-9062-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICPR56361.2022.9956530

***Abstract*—Detection and segmentation of surgical instruments is an important problem for laparoscopic surgery. Accurate pixel-wise instrument segmentation is a useful intermediate task for the development of computer-assisted surgery systems, such as pose estimation, surgical phase estimation, enhanced image fusion, video retrieval and others. In this paper we describe a deep learning-based approach to instrument segmentation, which addresses the binary segmentation problem in which every pixel in an image is labeled as instrument or background. The key novelty of our approach relates to the use of training data which is inexpensive and fast to acquire. First, our approach relies on weak annotations provided as bounding boxes of the instruments, which are much faster and cheaper to obtain than a dense pixel-level annotations. Second, to further improve the system's accuracy we propose a novel approach to generate synthetic training images. Our approach achieves state-of-the-art results, outperforming previously proposed methods for automatic instrument segmentation, based only on weak annotations.**

I. Introduction

Laparoscopic surgery has changed surgical practice by reducing operative trauma, visible scars and the hospitalization period. In such minimally invasive surgery (MIS), surgeons access the body through several small incisions and observe the internal anatomy using one or more cameras. Most interactions with the internal anatomy and organs are therefore recorded digitally. The availability of such visual data combined with the difficulty and steep learning curve of MIS motivates the pursuit of vision-based approaches for analysing laparoscopic videos and the development of computer-assisted interventions (CAI) systems. The detection and segmentation of laparoscopic instruments, the tools used by the surgeon during the MIS, is an intermediate task in the development of various computer vision algorithms. These include, amongst others: surgical workflow analysis [1], surgical safety [2], [3], and surgeon skill assessment [4], [5]. These algorithms form the core for surgical assistance systems that range from improving the surgery's outcome, through operating room optimization, to surgeon training. Hence, developing reliable methods for surgical instrument detection and segmentation has the potential to advance multiple fields of research.

Instrument segmentation can be treated as a binary or instance segmentation problem for which classical ML algorithms have been applied using color and/or texture features [6]. Others formulated this problem as semantic segmentation, aiming at distinguishing between different instruments [7]. Recently, deep learning-based approaches have demonstrated superior performance over conventional ML methods for many problems in general computer vision domain [8], [9] as well as for the medical domain [10], [11]. Previous deep learning-based applications to instrument segmentation have demonstrated competitive performance in binary segmentation [12], [13] and in multi-class segmentation [14].

When using supervised deep learning-based approaches for solving instrument segmentation, a limiting factor is the relative scarcity of annotated data. Supervision is usually provided as bounding box coordinates for tool detection and as pixel-level annotations for tool segmentation. Bounding box annotations are relatively easy to collect compared with pixel-level annotations. However, previously proposed detection-style methods which use this form of annotation directly have difficulty in localizing the tools precisely; this is due to the fact that the bounding boxes usually include a large portion of the background. Therefore, most of the existing methods rely on pixel-level annotations, which are very time-consuming and expensive to collect at scale. In our experience, annotating a single image with a bounding box is much faster than pixel-level annotation (generally seconds vs. minutes).

In this paper we propose a novel approach to deal with the paucity of training data. Our method relies on weak and fast annotations provided as bounding boxes of the instruments. This data is further augmented with novel synthetic training images. Our approach achieves state-of-the-art results, outperforming previously proposed methods for automatic instrument segmentation, based on weak annotations only.

The main contributions of this paper are threefold:

1) We introduce a novel approach for instrument segmentation in laparoscopic surgeries that relies on weak and fast annotations provided as bounding boxes.
2) We propose to incorporate synthetic images into the training. The generation of the synthetic images is done by rendering laparoscopic scenes using Blender3D [15], and then making these images more realistic by applying CycleGAN.
3) We show that training on the combination of automatically generated segmentation with our synthetic images outperforms previously proposed methods.

The remainder of the paper is organized as follows. Section II provides a review of the related work. Section III describes the methods we propose. Section IV describes the datasets we use. The results and their discussion are presented in Section V. Finally, Section VI concludes the paper.

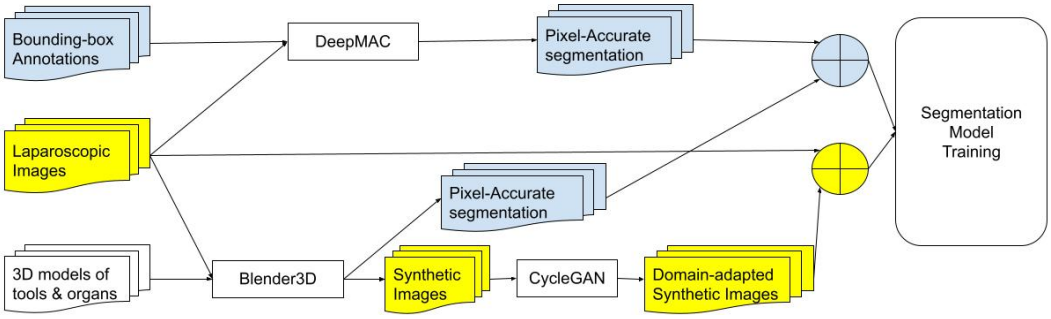


Fig. 1. **Overview of our approach:** Given a dataset of laparoscopic images and their corresponding bounding box annotations, we generate pixel-accurate segmentations using DeepMAC. Synthetic images are generated by Blender3D and are then passed through CycleGAN which performs domain adaptation. To train the segmentation model, the above types of data are combined. Images are in yellow, annotation and automatic masks are in blue. The crossed circles combine the read data with the synthetic one.

II. Related Work

An early comparative study of vision-based methods for instrument segmentation in laparoscopic surgery was presented in [16]. The segmentation methods described in this paper are based on either on random forests or relatively old CNNs. For example, [17] uses Random Forest classifier to distinguish instrument pixels from background pixels in a feature space where each pixel is represented by values from multiple color spaces and by gradient information. The dataset they used is relatively small compared to [18].

The largest comparative study on surgical instrument segmentation conducted recently is described in [18]. They organized the Robust Medical Instrument Segmentation (ROBUST-MIS) challenge. They described ten different methods that participated in the challenge, ranging from 2D U-Net variants (TernausNet [19], multi scale U-Net [20]) to different implementations of the Mask R-CNN [21] with a ResNet backbone to the DeepLabV3 [22] network architecture. We use the dataset in this paper to evaluate our approach and show that we outperform their best performing method.

In [23] the authors present an attention-pruned multi-task learning model (AP-MTL) by optimizing the task-aware MTL model to obtain the same optimal convergence point for multi-tasks. They use skip-scSE [24] to reduce sparsity and redundancy in the decoder. The evaluation provided in the paper is on the EndoVis2017 dataset [25].

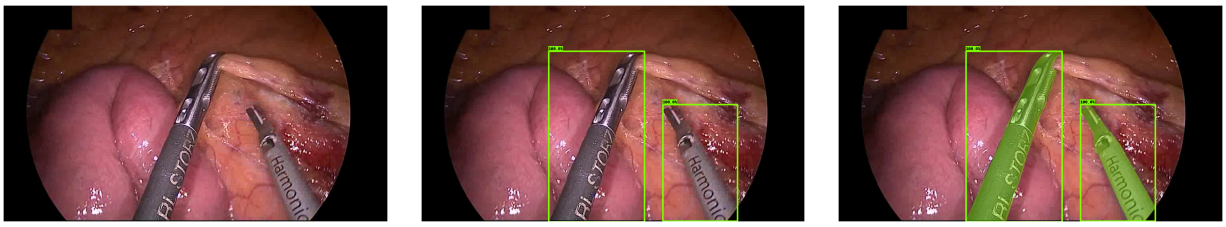
In [26] the proposed approach relies on weak annotations provided as stripes over the objects in the image. Then partial cross-entropy is used as the loss function of a fully convolutional neural network to obtain a dense pixel-level prediction map. While the proposed stripes can be more intuitive for a novice annotator, we have witnessed that the experienced annotators we work with prefer bounding-box annotation. Moreover, bounding boxes annotation is common practice, leading to availability of numerous pre-trained models. The stripes approach is limited to representing the instruments which are rigid, while it is harder to extend to generic surgical objects that can be non-rigid like specimen-bags.

In [27] the authors propose a method for segmentation of surgical tools without any manual annotations. They automatically generate labels for surgical tool segmentation by performing a semi-synthetic blending. Their approach is not end-to-end, as it requires post-processing. Moreover, the process of acquiring the semi-synthetic data is time-consuming since foreground data is needed to be captured by placing sample surgical instruments over a chroma key (a.k.a. green screen) in a controlled environment. While their results are impressive, they have not evaluated their approach on the most challenging datasets like EndoVis2019 [18]. In [28] the authors use CycleGAN [29] to correct automatically generated segmentations of robotic instruments produced by inverse kinematics. The training data is produced automatically but the results of their method is inferior compared to the accuracy of fully supervised algorithms.

Recently, to alleviate the need for annotation, simulation methods were proposed to produce synthetic data [30]–[33]. Simulation data can generate reliable training labels but has significant limitations in reproducing realistic features. To transfer the realistic style from real surgical frames to simulated ones, bridging the domain gap between simulation and real endoscopic images, image-to-image translation (I2I) is employed for transfer of style features from different image domains without the need for paired-samples. The works of [34], [35] also bridge the domain gap between simulation and real endoscopic images, by using joint and teacher–student learning approaches that learn from annotated simulation data. The results obtained are visually impressive, though the accuracy of the proposed methods is still lower than the accuracy of the segmentation models trained on real data.

III. Methods

The overview of our approach is demonstrated in Figure 1. Given a dataset of laparoscopic images and their corresponding bounding box annotations, we automatically generate pixel-accurate segmentations using the DeepMAC [36] (see Section III-A). In parallel, we generate synthetic training images by rendering scenes using Blender3D (see Section III-B). To bridge the gap between the generated synthetic images and real laparoscopic domain we use CycleGAN (see Section III-C).



(a) (b) (c)

Fig. 2. **Automatic segmentation:** Given image (a), we annotate a bounding box around each instrument (b). Then we automatically generate the segmentation for each instrument based on the bounding box (c) using DeepMAC.

Finally, to train the segmentation model, which will be used for inference, the images and their segmentations generated by the DeepMAC are combined with the synthetic images with their pixel-accurate segmentations into a single training set.

A. Automatic Segmentation Generation

Our approach relies on weak and fast annotations provided as bounding boxes of the instruments. Given an image (Fig. 2a), we annotate a bounding box around each instrument (Fig. 2b). Then we automatically generate segmentation for each instrument based on the bounding box (Fig. 2c). This accurate automatic generation of the segmentation is possible due to the recently proposed method known as DeepMAC [36]. The method builds instance segmentation capabilities on top of CenterNet [37], a popular anchor-free *detection* deep network that models each object by the center of its bounding box. In order to predict bounding boxes it outputs 3 tensors:

1) A per-class heatmap which indicates the probability of the bounding-box center being present at each location.
2) A class-agnostic 2-channel tensor indicating the height and width of the bounding box at each center pixel.
3) Since the output feature map is typically smaller than the image, CenterNet also predicts x- and y-offsets to recover this discretization error at each center pixel.

To extend the approach to *segmentation*, as explained in [36], a fourth pixel embedding branch P is added in parallel to the box-related prediction heads. For each bounding box b, a region P_b is cropped from P corresponding to b via ROIAlign [21], which results in a 32×32 tensor. Each P_b is fed to a second stage mask-head network, which is based on the Hourglass architecture [38]. The final prediction is a class-agnostic 32×32 tensor which is post-processed into a binary mask at test time by applying a sigmoid and thresholding at 0.5. This mask-head is trained via a per-pixel sigmoid cross-entropy loss averaged over all pixels and instances.

For automatic generation of the pixel-accurate segmentations from the given bounding boxes of the surgical instruments we use the above DeepMAC network as implemented in [36], which has been trained on the COCO dataset [39]. In this paper we assume that the only ground-truth available for laparoscopic scenes is provided as bounding boxes only and no masks are available. Therefore, we run DeepMAC inference that receives as input both an image as well as its corresponding bounding-box(es). Due to the presence of the bounding box, the first stage of the network may be skipped. Instead, the inputs may be fed directly into the second stage mask-head network, thereby producing segmentation mask(s) corresponding to the input bounding-box(es). Interestingly, we note that no additional training of DeepMAC on surgical tools is required; rather, we use network as pre-trained on the COCO dataset only. Empirically, we see that this is sufficient to produce reasonably quality segmentations of the tools corresponding to their bounding boxes.

After generating pixel-accurate segmentations for the training examples, we train another segmentation network, which is based on CenterNet architecture with the fourth pixel embedding branch, as explained above. The new network is trained using the ground-truth, which includes both bounding boxes and pixel-accurate segmentation.

As demonstrated in Figure 2(c) usually the automatic segmentation is very accurate. However, in some cases, specifically the ones where the bounding boxes of the instruments are overlapping, the segmentation is not perfect. An example of such case can be seen in Figure 3. To overcome this problem and to improve the overall accuracy even further we propose the following approach to generate synthetic training images.

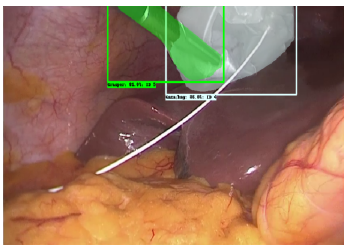
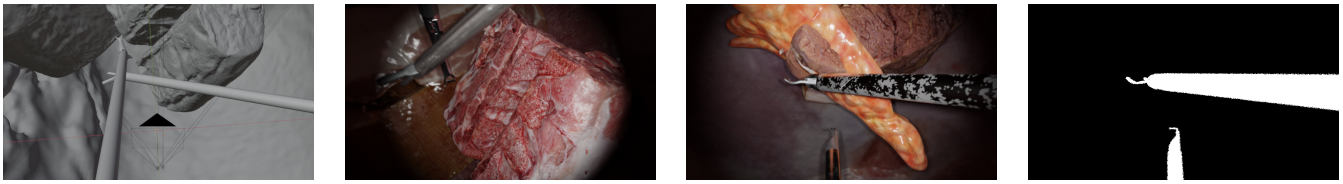


Fig. 3. **Limitations of Automatic Segmentation:** when the bounding boxes are overlapping the results of the automatically generated segmentation are not always satisfactory.

B. Synthetic Data Generation

Generation of annotations is an expensive process, especially for medical data as it requires special expertise. In contrast, annotations of synthetic data can be easily produced as part of the generation process (e.g. Fig. 4d). We use Blender3D [15] to render synthetic images of laparoscopic scenes. The scenes are made of 3D models of laparoscopic tools and 3D models that depict the background in an abdominal environment (e.g. Fig. 4a). An endless number of



(a) (b) (c) (d)

Fig. 4. **Synthetic dataset:** (a) A synthetic image is created by constructing a 3D scene of laparoscopic tools, 3D models representing abdominal organs as background, lighting and camera. (b) 3D rendering software renders the 3d scene from the camera view point. (c) Variability is created by randomizing tool and background types and position. Further variability is achieved by altering the texture of the tools. (d) Accurate segmentation masks can be generated by the same rendering tool.

different scenes can be generated by varying the combination, position and orientation of tools and background models; examples are shown in Fig. 4b and Fig. 4c. Further variability can be achieved by randomizing the texture of the tools and backgrounds. The lighting and camera parameters are set to match the specifications of typical endoscopes. To further increase the variability of the the synthetic dataset, we use frames from real laparoscopy procedures as background to the synthetic tools. We extract the background frames from the *Cholec80* [1] dataset. The *Cholec80* dataset has per frame tool presence annotations. We use these annotations to select frames without any tools present. This way we guarantee that the only tools in the rendered images are the synthetic tools. The large number of combinations of tools, background models and lighting conditions lead to enables large variability.

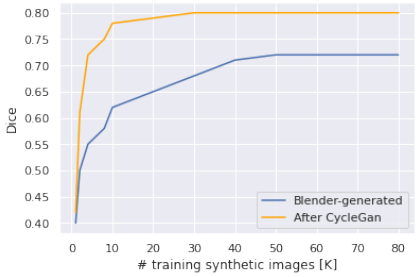
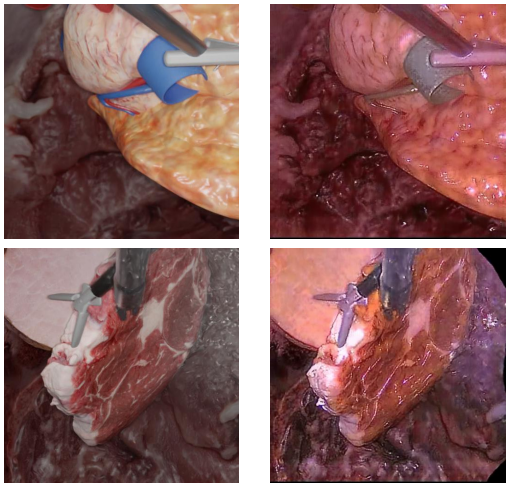


Fig. 5. **Synthetic Data Contribution:** The contribution of synthetic data to the segmentation performance increases with size of the synthetic dataset. The blue curve: without the use of CycleGAN, the orange curve: with CycleGAN. (Results shown for Endovis2019, Stage 3, Dice criterion.)

C. Domain Adaptation

Our approach for generation of synthetic images aims to make the 3D scene as close as possible to real laparoscopic images. Nevertheless, there remains a gap between the rendered and the real images. This is evident in the low accuracy achieved by training only on the 3D rendered dataset (blue plot in Fig. 5). To close this gap we train CycleGan [40] to perform domain-adaptation between the 3D rendered images and real laparoscopic images. We transform the 3D rendered dataset using the trained CycleGAN model to create the final synthetic dataset, which consists of images resembling real laparoscopic domain (Fig. 6). Both the accuracy and the convergence speed of the new dataset is superior to that of the 3D rendered only dataset (orange vs. blue in Fig. 5).



before CycleGAN after CycleGAN

Fig. 6. **CycleGAN dataset:** Examples of applying CycleGAN to synthetic images. The CycleGAN model emphasizes specular reflections on the tools and organs, modifies the scene colors, and rounds the image corners.

IV. Datasets and Evaluation Protocols

A. Public Datasets

The task of instrument segmentation in surgical scenes was first introduced in the Endoscopic Vision 2015 Instrument Segmentation and Tracking Dataset. However, the objective was not to distinguish among instrument types, but to extract the instruments from the background. The dataset's annotations were obtained using a semi-automatic method, leading to a misalignment between the ground truth and the images. Another limitation of this effort was the absence of substantial background changes, which further simplified the task.

The Endoscopic Vision 2017 Robotic Instrument Segmentation (EndoVis2017) Dataset [41] was developed to overcome the drawbacks of the 2015 benchmark. This dataset contains 10 robotic-assisted surgery image sequences, each composed of 225 frames. Eight sequences make up the training data and two sequences the testing data. Despite the effort put into building this dataset, it still does not reflect the general problem, mainly due to the limited amount of data.

To remedy the above issues EndoVis2019 [18] was presented at MICCAI 2019, containing more than 10,000 image frames for instrument segmentation. The dataset consists of

three stages with increasing difficulty in terms of their test sets. In Stage 1, the test data was taken from the procedures (patients) from which the training data were extracted. In Stage 2, the test data was taken from precisely the same type of surgery as the training data but from procedures (patients) not included in the training. And finally, in Stage 3 the test data was taken from a different but similar type of surgery (and different patients) compared to the training data.

Another real clinical dataset, called RoboTool [27] was released recently. This dataset contains 514 manually annotated images extracted from the videos of 20 freely available surgical procedures. The authors claim that the images contain many challenging scenarios with tool-tissue interaction, smoke, blood, debris and shadows. There is no train-test split provided; the dataset is used either for training or for testing.

B. Synthetic Dataset

We render our synthetic dataset using Cycles, Blender3D's ray-tracing engine, at a resolution of 855 $\times$ 480 pixels. We use 20 different models for background, 15 different models of laparoscopic tools and 10K frames taken from laparoscopic videos (where no tool is present). For each scene we randomly select two background models, or one background model and one real background frame, and one or two random laparoscopic tools. The background model and tools are randomly rotated and positioned allowing for occlusions. In total we have rendered 100K synthetic images using this method.

C. Evaluation

To assess the accuracy of segmentation the following two evaluation metrics are usually used: Jaccard index and Dice Similarity Coefficient. We use both methods, subject to the paper we compare to. The Jaccard index can be interpreted as a similarity measure between a finite number of sets. For two sets A and B, it can be defined as following:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

Since an image consists of pixels, the last expression can be adapted to images as follows:

$$J(A,B) = \frac{\sum_{i=1}^{|A|} a_i b_i}{\sum_{i=1}^{|A|} (a_i + b_i - a_i b_i)} \tag{2}$$

where A and B are images of the same size and a_i, b_i are the binary values (labels) of each pixel (0-background, 1-object). If A is the ground truth and B is the prediction, then a soft version of the above substitutes the predicted probability of the object for b_i. As is common, we use this soft version.

The Dice Similarity Coefficient is defined as the harmonic mean of precision and recall:

$$D(A,B) = \frac{2|A \cap B|}{|A| + |B|} \tag{3}$$

The following relationship between the Dice Similarity Coefficient (D) and the Jaccard index (J) can be easily shown:

$$D = \frac{2J}{J+1} \quad \Leftrightarrow \quad J = \frac{D}{2-D} \tag{4}$$

V. Results and Discussion

A. Quantitative comparison to SOTA

We compare our results to SOTA on three datasets: EndoVis2017 [25], EndoVis2019 [18] and RoboTool [27]. See Section IV-A for more details. To achieve an accurate and fair comparison, for each dataset we follow exactly the same setup and protocol as the paper we compare to.

1) EndoVis2017 dataset: Following the setup in [25] and [42] we use 4-fold cross-validation. The results are presented in Table I. The segmentation accuracy achieved by our model trained on the pixel-accurate annotations (D=0.931) is higher than [25] (D=0.900). Adding our synthetic data into the training improves the accuracy even further (D=0.949).

When no pixel-accurate segmentation annotations is provided, we automatically generate the segmentation for training data as explained in Section III-A. The accuracy of this model is only slightly lower ($D = 0.898$ vs. $D = 0.900$) than [25] which used the full pixel-accurate annotation in training. Furthermore, incorporating our synthetic data into training significantly improves the results, achieving D=0.941.

TABLE I
Segmentation results on EndoVis 2017 dataset

Training			Testing
Method	Annotation	Synthetic Data	(Dice)
RF [17]	pixel-accurate	-	0.841
AP-MTL [23]	pixel-accurate	-	0.871
Shvets et al. 2018 [25]	pixel-accurate	-	0.900
RASNet 2019 [43]	pixel-accurate	-	0.946
Our	pixel-accurate	-	0.931
Our	pixel-accurate	5000	0.949
Our	-	5000	0.881
Our	bounding box	-	0.898
Our	bounding box	5000	0.941

2) EndoVis2019 dataset: Following the setup in [18] we evaluate the results on three different stages with increasing difficulty. The results are presented in Table II. The segmentation accuracy achieved by our model trained on the pixel-accurate annotations is slightly lower than SOTA [18]. Adding our synthetic data into the training improves the accuracy and closes the above gap for all the stages.

TABLE II
Segmentation results on EndoVis 2019 dataset

Training			Testing (Dice)		
Method	Annotation	Synth. Data	Stage 1	Stage 2	Stage 3
RF [17]	pixel-accurate	-	0.84	0.82	0.82
AP-MTL [23]	pixel-accurate	-	0.87	0.85	0.84
Or-unet 2020 [18]	pixel-accurate	-	0.92	0.90	0.88
Our	pixel-accurate	-	0.90	0.88	0.87
Our	pixel-accurate	5000	0.92	0.91	0.89
Our	-	5000	0.74	0.76	0.80
Our	bounding box	-	0.88	0.89	0.85
Our	bounding box	5000	0.91	0.90	0.89

3) RoboTool dataset: Following the setup in [27] we test on the RoboTool dataset while training is performed on the EndoVis2017 dataset. The results are presented in Table III. The segmentation accuracy achieved by our model trained on the pixel-accurate annotations is slightly lower (0.685) than

SOTA [27] (0.694). Adding our synthetic data into the training improves the accuracy to 0.722, outperforming SOTA [27],

Note, that the numbers in Table III for the SOTA [27] are associated with applying their method in combination with an independent post-processing step based on GrabCut [44]. Without the application of GrabCut, their results are much lower: 0.666 (vs. 0.694) and 0.561 (vs. 0.681).

TABLE III
SEGMENTATION RESULTS ON ROBOTOOL DATASET.

Training			Testing
Method	Annotation	Synthetic Data	(IoU)
RF [17]	EndoVis17 pix-accurate	-	0.614
Garcia 2021 [27]	EndoVis17 pix-accurate	-	0.694
Garcia 2021 [27]	-	Semi-synth, 100000	0.681
Our	-	Our, 5000	0.67
Our	EndoVis17 pix-accurate	-	0.685
Our	EndoVis17 pix-accurate	Our, 5000	0.722
Our	EndoVis17 bounding-box	-	0.652
Our	EndoVis17 bounding-box	Our, 5000	0.711

B. Ablation and Sensitivity Studies

In the previous section we have provided quantitative results for various datasets. In this section, we provide ablation and sensitivity studies. To make the studies easier to follow, we concentrate on only one dataset, EndoVis2019 [18]. For training 5,983 images are used. For evaluation we use the EndoVis2019 - Stage 3 test set consisting of 2,880 images, which represents the largest of the all the test sets.

Figure 5 demonstrates the accuracy contribution of the synthetic data as a function of the size of the synthetic dataset. Using the images that underwent the CycleGAN domain-adaptation increases the accuracy of the segmentation by 10% compared to using the Blender-generated images directly. Moreover, 5000 CycleGAN synthetic training images are enough to achieve accuracy saturation; while when using Blender-generated images for training (without CycleGAN), we need around 30K images to achieve saturation.

Figure 7 shows segmentation performance vs. overall training set size (of real images). As shown in the previous section, highest accuracy is achieved when incorporating our synthetic data into the training. When using synthetic data, the contribution of the pixel-accurate annotations (blue) is almost identical to the case of using the cheap bounding box annotations (orange). We also see that when our synthetic data is used, we can reduce the number of real training images by half with only a very small decrease in the final accuracy. Note that in each of the experiments we used 5000 synthetic images; increasing this number does not increase the final accuracy.

Figure 8 shows the segmentation performance as a function of the percentage of pixel-accurate annotations. For each the data points on the curve we train on 6K real images from the EndoVis 2019 train set. The only thing that varies between the data points is the percentage of pixel-accurate annotation vs. bounding box annotation. The leftmost data point, 0%, corresponds to the case when all of the annotations provided during the training are bounding boxes. By contrast, the rightmost data point, 100%, corresponds to the case when

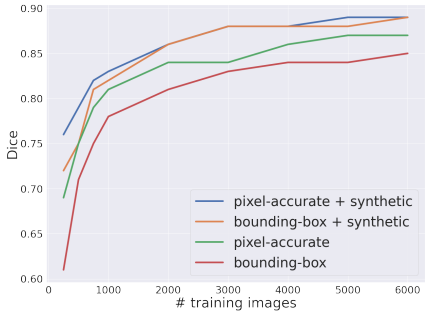


Fig. 7. **Sensitivity to the number of training images:** the segmentation accuracy increases as the number of the training images increases (Endovis2019, Stage 3, Dice criterion). See accompanying description in the text.

all of the annotations are pixel-accurate segmentations. It can be seen that using only 30% of pixel-accurate annotations already achieves the maximal accuracy (equivalent to using 100% pixel-accurate annotations). This indicates that we can significantly reduce the annotation effort without compromising accuracy, even before the introduction of the synthetic data.

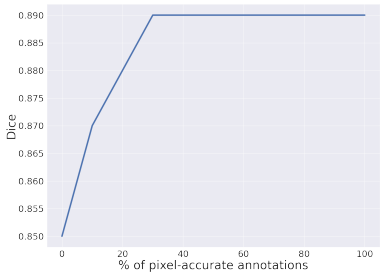


Fig. 8. **Sensitivity to percentage of pixel-accurate annotations:** each of the data points correspond to training on 6K real images, varying between the percentage of the pixel-accurate annotation vs. the percentage of bounding box annotation. (Endovis2019, Stage 3, Dice criterion)

VI. CONCLUSION

We introduced a novel approach for instrument segmentation in laparoscopic surgeries that relies on weak and fast annotations provided as bounding boxes of the instruments. To achieve an accurate automatic generation of the segmentation based on bounding boxes only we employed DeepMAC [36]. To improve the accuracy we incorporated synthetic images into training. The generation of the synthetic images is achieved by rendering laparoscopic scenes using Blender3D. To bridge the gap between the generated synthetic images and the real laparoscopic domain we used CycleGAN [29].

Our approach achieved SOTA results based on weak annotations only, Training on the combination of automatically generated segmentations with our synthetic images, outperformed previously proposed methods for instrument segmentation. In the future we plan to extend our work from binary segmentation to instrument part segmentation. To improve the accuracy we plan to introduce ensemble diversification methods [45]. We also plan to explore how our synthetic data can enable the 3D pose estimation of the instruments.

REFERENCES

[1] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, "Endonet: a deep architecture for recognition tasks on laparoscopic videos," *IEEE transactions on medical imaging*, vol. 36, no. 1, pp. 86–97, 2016.

[2] V. Gupta and G. Jain, "The r4u planes for the zonal demarcation for safe laparoscopic cholecystectomy," *World Journal of Surgery*, vol. 45, no. 4, pp. 1096–1101, 2021.

[3] ——, "Safe laparoscopic cholecystectomy: Adoption of universal culture of safety in cholecystectomy," *World Journal of Gastrointestinal Surgery*, vol. 11, pp. 62–84, 2019.

[4] A. Jin, S. Yeung, J. Jopling, J. Krause, D. Azagury, A. Milstein, and L. Fei-Fei, "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, 2018, pp. 691–699.

[5] S. Lin, F. Qin, R. A. Bly, K. S. Moe, and B. Hannaford, "Automatic sinus surgery skill assessment based on instrument segmentation and tracking in endoscopic video," in *International Workshop on Multiscale Multimodal Medical Imaging*. Springer, 2019, pp. 93–100.

[6] C. Doignon, F. Nageotte, and M. De Mathelin, "Segmentation and guidance of multiple rigid objects for intra-operative endoscopic vision," in *Dynamical Vision*. Springer, 2006, pp. 314–327.

[7] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting surgical tools by modelling local appearance and global shape," *IEEE transactions on medical imaging*, vol. 34, no. 12, pp. 2603–2617, 2015.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[9] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[10] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *Medical image analysis*, vol. 35, pp. 18–31, 2017.

[11] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[12] L. C. Garcia-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5717–5722.

[13] M. Attia, M. Hossny, S. Nahavandi, and H. Asadi, "Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2017, pp. 3373–3378.

[14] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2019, pp. 566–573.

[15] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[16] S. Bodenstedt, M. Allan, A. Agustinos, X. Du, L. Garcia-Peraza-Herrera, H. Kenngott, T. Kurmann, B. Müller-Stich, S. Ourselin, D. Pakhomov *et al.*, "Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery," *arXiv preprint arXiv:1805.02475*, 2018.

[17] S. Bodenstedt, M. Wagner, B. Mayer, K. Stemmer, H. Kenngott, B. Müller-Stich, R. Dillmann, and S. Speidel, "Image-based laparoscopic bowel measurement," *International journal of computer assisted radiology and surgery*, vol. 11, no. 3, pp. 407–419, 2016.

[18] T. Roß, A. Reinke, P. M. Full, M. Wagner, H. Kenngott, M. Apitz, H. Hempe, D. Mindroc-Filimon, P. Scholz, T. N. Tran *et al.*, "Comparative validation of multi-instance instrument segmentation in endoscopy: Results of the robust-mis 2019 challenge," *Medical image analysis*, vol. 70, p. 101920, 2021.

[19] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[20] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[23] M. Islam, V. Vibashan, and H. Ren, "Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8433–8439.

[24] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 421–429.

[25] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 624–628.

[26] F. Fuentes-Hurtado, A. Kadkhodamohammadi, E. Flouty, S. Barbarisi, I. Luengo, and D. Stoyanov, "Easylabels: weak labels for scene segmentation in laparoscopic videos," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1247–1257, 2019.

[27] L. C. Garcia-Peraza-Herrera, L. Fidon, C. D'Ettorre, D. Stoyanov, T. Vercauteren, and S. Ourselin, "Image compositing for segmentation of surgical tools without manual annotations," *IEEE Transactions on Medical Imaging*, 2021.

[28] D. Pakhomov, W. Shen, and N. Navab, "Towards unsupervised learning for instrument segmentation in robotic surgery with cycle-consistent adversarial networks," *arXiv preprint arXiv:2007.04505*, 2020.

[29] S. Chen and C. Xia, "Cyclegan, https://github.com/tensorflow/gan," 2020.

[30] M. Pfeiffer, I. Funke, M. R. Robu, S. Bodenstedt, L. Strenger, S. Engelhardt, T. Roß, M. J. Clarkson, K. Gurusamy, B. R. Davidson *et al.*, "Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 119–127.

[31] K. Lee, M.-K. Choi, and H. Jung, "Davincigan: Unpaired surgical instrument translation for data augmentation," in *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, pp. 326–336.

[32] E. Colleoni and D. Stoyanov, "Robotic instrument segmentation with image-to-image translation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 935–942, 2021.

[33] E. Colleoni, P. Edwards, and D. Stoyanov, "Synthetic and real inputs for tool segmentation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 700–710.

[34] M. Sahu, R. Strömsdörfer, A. Mukhopadhyay, and S. Zachow, "Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 784–794.

[35] M. Sahu, A. Mukhopadhyay, and S. Zachow, "Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation," *International journal of computer assisted radiology and surgery*, pp. 1–11, 2021.

[36] V. Birodkar, Z. Lu, S. Li, V. Rathod, and J. Huang, "The surprising impact of mask-head architecture on novel class segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7015–7025.

[37] X. Zhou, D. Wang, and P. Krahenb, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[38] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*. Springer, 2016, pp. 483–499.

[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[40] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[41] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[42] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: An instance-based approach for surgical instrument segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 595–605.

[43] Z.-L. Ni, G.-B. Bian, X.-L. Xie, Z.-G. Hou, X.-H. Zhou, and Y.-J. Zhou, "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)*. IEEE, 2019, pp. 5735–5738.

[44] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," *ACM transactions on graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[45] M. Farber, R. Goldenberg, G. Leifman, and G. Novich, "Novel ensemble diversification methods for open-set scenarios," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1485–1494.