Noname manuscript No.
(will be inserted by the editor)

EasyLabels: Weak labels for scene segmentation in laparoscopic videos

Félix Fuentes-Hurtado · Abdolrahim Kadkhodamohammadi · Evangello Flouty · Santiago Barbarisi · Imanol Luengo · Danail Stoyanov

Received: date / Accepted: date

Abstract

Purpose We present a different approach for annotating laparoscopic images for segmentation in a weak fashion and experimentally prove that its accuracy when trained with partial cross-entropy is close to that obtained with fully-supervised approaches.

Methods We propose an approach that relies on weak annotations provided as stripes over the different objects in the image and partial cross-entropy as the loss function of a fully convolutional neural network to obtain a dense pixel-level prediction map.

Results We validate our method on three different datasets, providing qualitative results for all of them and quantitative results for two of them. The experiments show that our approach is able to obtain at least 90% of the accuracy obtained with fully-supervised methods for all the tested datasets, while requiring $\sim$ 13x less time to create the annotations compared to full supervision.

Conclusion With this work we demonstrate that laparoscopic data can be segmented using very few annotated data while maintaining levels of accuracy comparable to those obtained with full supervision.

Keywords Computer assisted interventions, laparoscopy, instrument detection and segmentation

1 Introduction

Laparoscopic surgery has changed surgical practice by reducing operative trauma, risk of co-morbidity, visible scars and hospitalization period. In such minimally

Félix Fuentes-Hurtado[1] · Abdolrahim Kadkhodamohammadi[1] · Evangello Flouty[1] · Santiago Barbarisi[1] · Imanol Luengo[1] · Danail Stoyanov[1,2]
E-mail: ffuentes@upv.es
[1]Digital Surgery. London, United Kingdom
[2]Wellcome / ESPRC Centre for Interventional and Surgical Sciences. London, United Kingdom

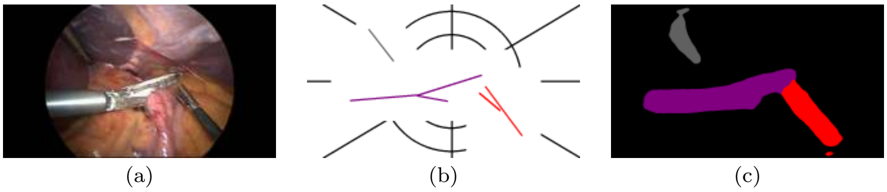

Fig. 1 Our approach to perform semantic segmentation with weak labels. (a) shows the original image, (b) the proposed annotations and (c) the full pixel-level segmentation mask obtained. We label our data using straight lines along the longitudinal axis of the instruments and train our network with partial cross-entropy as loss, obtaining very similar results to full supervision while drastically reducing the density of the annotations needed.

invasive surgery (MIS), surgeons access the body through several small incisions and observe the internal anatomy using cameras. Most interactions with the internal anatomy and organs are therefore recorded digitally. The availability of such visual data together with the necessity to enhance surgeons' capabilities to meet the difficulty and complexity of MIS (hand-eye coordination, restricted mobility and narrow field of view [2]) has driven computer assisted interventions (CAI) and computer vision based approaches to analyse laparoscopic video [19]. The localization and segmentation of surgical instruments (or even anatomy) are a prerequisite for many potential CAI applications ranging from intra-operative assistive systems for better navigation or surgical robotics to enhanced image fusion or video retrieval systems [2, 17].

Recent vision-based methods for tool localization and segmentation in images tend to take a supervised approach [2–4, 8–10, 13, 15, 18] while historical techniques are model based and not data driven [4]. Supervision is usually provided as bounding box coordinates for tool localization and as pixel-level annotations for tool segmentation. While bounding box annotations are relatively easy to collect compared with pixel-level annotations, they normally include a large portion of the background. As a result, it is not feasible to localise the tools precisely. On the other hand, providing pixel-level annotations enables models to localise tools accurately. However, the process to create such fine-level annotations is very time-consuming and tedious or expensive if implemented at scale. To alleviate these problems, we propose to localise tools using what we define as weak "stripe" annotations (Figure 1). We use this type of annotation for two main reasons: (1) it is as quick to do as bounding box annotations and (2) it represents the geometric nature of surgical instruments better, as most of them are mostly rigid and highly linear.

2 Related Work

Surgical instrument localization and segmentation are extremely difficult tasks due to background variability and challenges such as smoke, blood, visual occlusions, shadows and specular reflections. These effects and challenges are often present in laparoscopic videos. In the literature, most attempts to perform either localization or segmentation of tools are based on fully supervised methods (i.e. with pixel-level annotations) [4] using Fully-Convolutional Neural Networks (FCNN), the *de* *facto* choice for approaching this kind of problem, not just in CAI, but in general computer vision.

For example, Garcia-Peraza-Herrera et al. [9] combine a FCNN and optical flow to accomplish real-time segmentation and tracking of non-rigid surgical tools. Pakhomov et al. [15] develop deep neural networks with residual connections to segment instruments in robotic surgery. Laina et al. [10] leverage FCNNs to perform concurrent segmentation and localization of surgical instruments. Shvets et al. [18] propose the combination of different FCNNs to tackle both binary and multi-class segmentation of robotic instruments. Finally, some works still rely on traditional methods, such as the one of Bondenstedt at al. [3], which makes use of Random Forests and features such as histograms over hue and saturations, gradients or SURF features, to achieve real-time instrument classification for laparoscopic surgery. Nevertheless, they often face the same problem: the lack of annotated data due to the difficulty of annotating with pixel-level labels. Most of the available datasets consist only of a few thousand images with a maximum of 5-6 sequences [4]. This is a limiting factor on the research of new methods for segmentation or localization.

For this reason, there has been a rise in the interest of finding methods that can overcome this limitation (i.e. achieve accurate segmentations without full annotations). There are two different ways to address the segmentation with weak labels: one consists in artificially creating "fake" full segmentation masks ("proposals") and then treating the problem as a fully supervised one (minimizing the cross-entropy w.r.t such "proposals"/masks) [21]; whereas the other relies on using only small portions of the whole image or "scribbles"[1] to train the network while ignoring the rest of the image (i.e. minimizing the cross-entropy w.r.t. only the pixels within the "scribbles") [1,12,20,23,24]. The proposals approach often leads to worse results since the created "fake" masks are usually wrong, which reinforces the network to learn wrong patterns, and eventually makes it fail in producing accurate, correct segmentation masks [20]. In contrast, approaches using "scribbles" have recently shown greater accuracy and robustness, as they focus on making the network learn from less but still informative, correct information.

Some of these approaches are using points [1,11], scribbles [12,20], bounding boxes [24] or even image labels [23]. However, to the best of our knowledge, none of them has yet been applied to surgical data for segmentation purposes. There exist some works trying to alleviate the burden of fully labeling datasets, such as the one of *Ross et al.* [16], who employ Generative-Adversarial Networks (GANs) to reduce the amount of labeled data needed for training. However, they still use full pixel-level segmentation masks. Vardazaryan et al. [22] implement a fully convolutional neural network and use class peak response for surgical instrument localization employing as ground truth only the image labels. However, this work is not yet able to obtain a segmentation mask of the surgical instruments.

There are some other different paths to try to alleviate this problem, such as crowd-sourcing or label propagation [7]. In the case of crowd-sourcing, it does not reduce the burden, just re-distributes it, and could also benefit from weak labels. Label propagation consists on fully annotating a single image and then propagate these labels, correcting them when needed. Although it increases the speed, it is

[1] For simplicity, even if there are approaches either using points, bounding boxes or scribbles, we will refer to this approach as just "scribbles".

still necessary to check if the propagated labels are correct and if they are not, to correct them, which leads us back to the same problem. For all these reasons, we propose an alternative to alleviate these caveats.

In summary, in this paper we introduce a fast method for annotating surgical datasets which can be used to perform semantic segmentation with results close to fully-supervised methods while requiring significantly less effort to create the annotations, as can be observed in Figure 1. To support it, we present an experimental evaluation on three different datasets showing the benefits of the introduced method for surgical segmentation tasks.

3 Methods

The proposed method takes an alternative approach for labeling images in a weak manner for CNN segmentation, combined with a slight modification of plain cross-entropy.

3.1 Stripes: our annotation method for weakly-supervised tasks

The proposed method consists in using straight rigid lines (i.e. defined by two points) to annotate the foreground objects along their longitudinal dimension (Figure 1 (b)). We call these lines “stripes”. There is no need to annotate the background, since it is automatically annotated using an artificially generated grid and the stripes corresponding to the foreground objects. By simplifying the annotation process, we are able to drastically reduce the time needed to annotate surgical data. These “stripes” along the foreground objects allow us to effectively train a network able to learn the patterns needed to perform accurate segmentation. As a rule of a thumb, lines must be close to boundaries among different classes at some point in order to give better accuracy in these regions. Indeed, by placing the lines as if they mimic the skeleton[2] of the blobs present in a hypothetical pixel-wise mask, our method encourages the network to focus on the most discriminating patterns so that it is able to generalise and segment the rest of the image in an accurate way. We employ two types of annotations: the automatic ones obtained by computing the skeleton of the full annotation masks and the manual ones performed by humans. Furthermore, we also test an additional reduced human annotations set to check how a decrease in the amount of labeled data affects the accuracy of the method.

3.1.1 Automatic weak mask generation for foreground objects from full annotation masks

We compute the distance-based skeleton on the foreground blobs to automatically obtain the stripes using a full annotation mask as input. As the background mask is automatically generated (see Section 3.1.3) we keep only the annotations corresponding to foreground objects. We call these annotations “auto stripes”.

[2] Please note that the skeleton is computed as the ridge of the distance transform.

3.1.2 Human weak mask generation for foreground objects

To prove that the computed skeleton is comparable to the annotations a human would do, we also annotate the Endovis15 dataset manually. In addition, to be able to assess how the amount of data affects this kind of annotations, we manually label this dataset in two different ways. On the one hand, we annotate the foreground objects using two stripes covering the different ends of the tool and crossing in the middle. This allows us to annotate both the center of the tool and the extremes at the same time, as shown in Figure 2 (c), first row. In case there is a hole in the shape, we use enough stripes to “round the hole”, as shown in Figure 2 (c), second row. We call these annotations “human stripes”. On the other hand, we create another set of annotations called “reduced human stripes” in which we keep the annotations as simple as possible, employing a single stripe per blob, except when holes are present, where we employ two lines (Figure 2, last column).

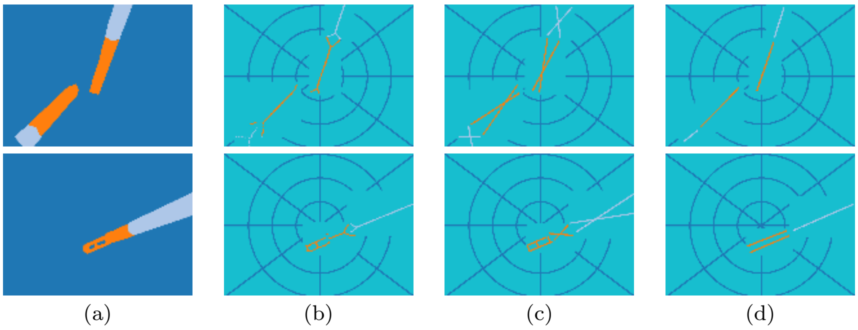

Fig. 2 Comparison of automatic and human annotations: (a) correspond to the full annotation mask, (b) to the automatically computed weak annotation mask, (c) to the manual annotations and (d) to the reduced set of manual annotations. Note that the background grid is always automatically computed using the foreground annotations.

3.1.3 Background grid generation

The background grid employed in this work consists of four lines forming a star and three concentric circles (Figure 3 (b)). However, this grid can be modified according to the needs of each dataset. Then, we employ the previously explained foreground stripes to automatically create a binary background/foreground mask that will remove the unnecessary pixels from the background grid. This binary mask excludes the regions covered by the bounding boxes obtained from the stripes (either automatically computed or manually labeled). Figure 3 (d) shows the result of applying the background/foreground mask to the background grid. To obtain the complete weak mask, we combine the foreground and background masks together and assign all the remaining pixels a value to indicate that we can ignore them when computing the loss, as shown by Figure 3 (e). This approach allows us to automatically label the background even when the full annotation mask is not available, and it further reduces the time needed per annotation, since it removes the necessity of manually labeling the background. In average, our weak labels account for less than 5% of the full annotation masks, while yielding promising segmentation results, as Section 6 shows.

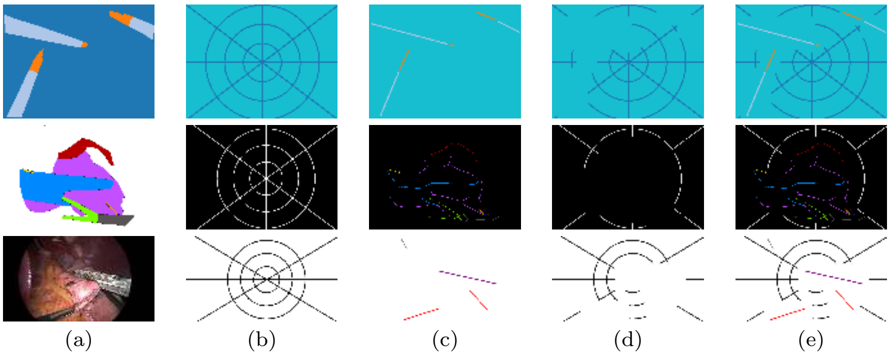

Fig. 3 Pipeline to automatically create our background weak annotations for each dataset: (a) shows the original full mask (except for Bypass, the dataset without full annotation masks available, where the original image is shown instead), (b) the automatically created background grid, (c) the foreground annotations (manual when available, automatic elsewhere), (d) the automatic background annotations obtained from (b) and (c) and (e) the final weak annotations mask.

3.2 Deep Learning architecture

The model employed in this work belongs to the family of the U-shaped Fully Convolutional Networks (FCN). These networks normally consist of an encoder, composed by several convolutional blocks that extract features from the input image while down-sampling the feature maps; and a decoder, that works similarly but up-sampling the feature maps while attempting to map the computed features to the correct classes. For this work, we use the *DeepLabv3+* [5] architecture, a FCN that consists of an encoder and a decoder. The encoder is based on the Xception network [6], which allows to build a rich representation of the input data while keeping the number of parameters low thanks to the depth-wise separable convolution. Then, the decoder makes use of atrous convolutions to retain more information from the boundaries, which yields better representations for the final segmentation mask. We trained DeepLabv3+ both using our ground truth and partial cross-entropy as loss, and using the full masks and normal cross-entropy when available. All the experiments were ran on two NVIDIA GTX 1080 Ti GPUs with 11GB of VRAM memory each. We chose the Stochastic Gradient Descent optimizer with a batch size of 10 and learning rate of $1e^{-4}$. The number of epochs was established at 100 for Endovis, and 50 for the other datasets. As our objective was to give a baseline of how accurate results it is possible to obtain by just using "stripes" and partial cross-entropy as loss, we did not attempt to optimize the hyper-parameters of the model.

3.3 Partial Cross Entropy as Loss Sampling

Cross-entropy loss is probably the gold standard when training neural networks for classification tasks. It computes the negative log-likelihood between the predictions and the true labels. In the case of *partial* cross-entropy, the same operation is performed but only for the *indicated* elements. Therefore, the only difference between plain cross-entropy and partial cross-entropy is the use of an indicator function that samples the loss to compute only those pixels within the weak annotations. It can be mathematically defined as:

$$\sum_{p \in \Omega} -u_p \cdot log S_p^{y_p} \tag{1}$$

where Ω is the image domain, $S_p \in [0,1]^K$ describes the network's output for $p \in \Omega$, y_p is the true label of p, K is the number of classes, and u_p is the indicator function, defined as $u_p = 1$ for $p \in \Omega_{\mathcal{L}}$ and 0 otherwise. $\mathcal{L}$ is the subset of the image domain Ω corresponding to the weak annotation pixels.

4 Datasets

We tested our method on three different datasets: two developed and labeled in-house, and the third one publicly available Endovis 2015 Rigid Instruments dataset [2, 14]. We describe them in detail in the following subsections.

4.1 Endovis 2015 Rigid Instruments dataset

This dataset consists of 40 2D in-vivo images for each of the 4 laparoscopic colorectal surgeries with their corresponding annotated masks for training (160 images), and 140 images for testing: 10 images from each of the 4 training videos and 50 images from 2 new videos. Two types of full masks are provided: one with the background, shaft and manipulator labeled; and another in which each tool instance on an image is annotated with a different label. For this work, we select the ground truth differentiating among background, shaft and manipulator. Since the available ground truth for this dataset consists of full masks, we followed the two different approaches previously explained (automatic and manual) to obtain the weak annotation masks. Furthermore, we perform an additional annotation pass reducing the amount of annotations, as explained in Section 3.1.2, to test how the amount of data affects our approach.

4.2 Sleeve Gastrectomy dataset

We tested our approach on another dataset consisting of 5 videos recorded during a laparoscopic sleeve gastrectomy totalling 3600 frames with their corresponding annotations. These annotations are full pixel-level segmentation masks differentiating among the different surgical instruments, anatomy and background. Concretely, this dataset consists of the following 14 labels for instruments and anatomy: stapler, stapler handle, stapler trigger, atraumatic grasper handle, atraumatic grasper tip, liver retractor, ligasure tip, ligasure handle, marylands tip, marylands handle, bandage, liver, stomach and background. Since the available annotations for this dataset are dense pixel-wise masks, we follow the procedure explained in Section 3.1 to automatically compute the weak masks. Figure 4 shows some examples of the original images, the full annotation masks and the computed weak masks.

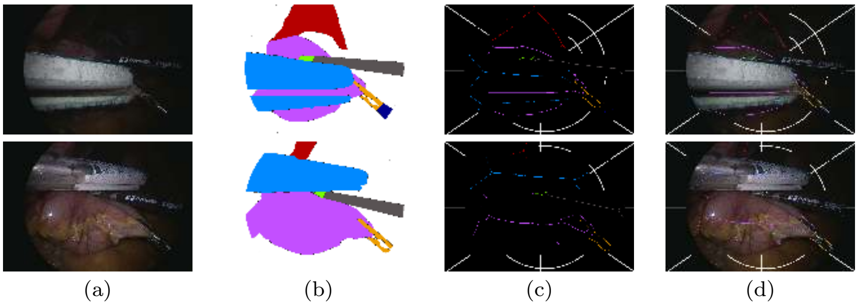

Fig. 4 Some sample images from our Sleeve Gastrectomy dataset (a) together with their corresponding pixel-level full annotation masks (b), the annotation mask we generated from the full mask to test our approach (c) and an overlay of the original image with our proposed ground truth (d).

4.3 Gastric Bypass dataset with "stripes"

Our other dataset consists of 20 videos of laparoscopic gastric bypass totalling 46606 frames with their corresponding *weak* annotations. These annotations were carried out in-house by informed participants using our "stripes" approach. Concretely, this dataset consists of 14 labels for the following instruments: atraumatic grasper, bowel clamp, clip applicator, drain, harmonic scalpel, hook, ligasure, liver retractor, marylands, needle holder, o'reilly, scissors, stapler and suction irrigation. Figure 5 shows two sample images with their corresponding proposed weak annotations.

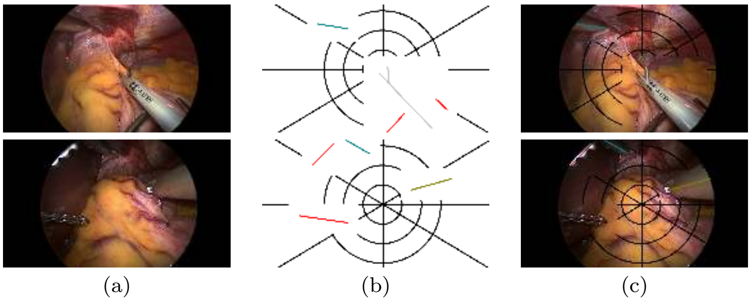

Fig. 5 Some sample images from our Gastric Bypass dataset. (a) shows the original images, (b) our proposed ground truth and (c) an overlay with the original image.

5 Validation studies

5.1 Endovis 15

We begin our validation studies with the Endovis 15 dataset using the same splits as proposed by the challenge in 2015 for its evaluation, that is, following a leave-one-video-out 5-fold cross-validation. The amount of weak annotations needed is an important matter to study so we can minimise the efforts of annotating data. Therefore, in addition to the full *versus* weak comparison, we perform an extra experiment with Endovis15 using a reduced set of annotations, as explained in Section 3.1.2. We provide results for each of the 5 folds and their average for each version of the dataset: full masks, automatically computed weak masks, manually annotated weak masks and reduced version of manually annotated weak masks.

5.2 Sleeve Gastrectomy dataset

To evaluate our approach on this dataset, we carry out two experiments: one with full masks and one with the automatically computed weak masks. We perform a 5-fold cross-validation following a leave-one-video-out fashion. This means that for each fold, four videos were used for training and one for evaluating. We evaluated our method for each fold using the automatically generated weak annotations and compared our results with those obtained with full segmentation masks.

5.3 Gastric Bypass dataset

Out of the 20 available videos, we employed 16 videos (37100 frames) for training and 4 (9506 frames) for testing. Since we do not have full pixel-level masks for this dataset, we only show qualitative results.

5.4 Time comparison

We measured the time taken by 5 in-house annotators for labeling a subset of 30 images of our Sleeve Gastrectomy dataset with pixel-level full annotations, and compared it to the time they needed to perform our "stripes" annotations on the same images. The time of each image was averaged for each annotator, and then all means were averaged to produce the final result. All of these images were non-consecutive frames, so the measures reported are for the worst case scenario. Commercial software was employed for the full annotations, while for the "stripes" annotations a software developed in-house was used.

6 Results

6.1 Endovis 2015 results

Table 1 presents the results obtained by our approach on the Endovis 2015 dataset with our annotations computed automatically, manually by a human, and with full supervision. It also shows the accuracy when the amount of annotations is reduced. Our method achieves almost 97% of the accuracy full supervision achieves. However, if the reduced annotations are used, there is a significant drop in the accuracy of the segmentation. Figure 6 shows some examples of the segmentation performed by our approach (with the annotations computed automatically) compared to the segmentation obtained with full supervision and the ground truth. In addition, Table 2 shows the amount of annotated pixels per class for each approach.

Table 1 Segmentation results for Endovis 2015. Per-class and averaged IOU values are reported for each fold. Folds indicate in which videos the model was evaluated. "weak human" and "weak human red." refer to the human annotations and the reduced set of human annotations (as explained in Section 3.1.2), respectively. Last column shows how similar the results are to those obtained with full supervision, from 0 to 1.

fold	mask	background	shaft	manipulator	avg	sim. (%)
OP5-OP6	weak human red.	.9957	.6934	.3585	.6172	.8783
	weak human	.9893	.8584	.5817	.7028	1
	weak auto	.9895	.8890	.5157	.7001	.9963
	full	.9947	.8868	.5325	.7027	-
OP4	weak human red.	.9857	.7126	.6833	.6929	.9309
	weak human	.9737	.8562	.7260	.6943	.9328
	weak auto	.9813	.8044	.7543	.7005	.9412
	full	.9908	.8704	.6662	.7443	-
OP3	weak human red.	.9928	.7906	.7173	.7652	.9298
	weak human	.9865	.8721	.8453	.7942	.9650
	weak auto	.9894	.8893	.7605	.7937	.9644
	full	.9969	.8539	.7870	.8230	-
OP2	weak human red.	.9892	.7986	.3750	.6618	.9036
	weak human	.9746	.9125	.5552	.6869	.9379
	weak auto	.9770	.6936	.5695	.6772	.9246
	full	.9898	.9060	.5139	.7324	-
OP1	weak human red.	.9972	.6521	.5467	.7022	.9129
	weak human	.9847	.8384	.8030	.7619	.9905
	weak auto	.9852	.8024	.7172	.7036	.9147
	full	.9960	.8196	.6687	.7692	-
avg	weak human red.	.9921	.7295	.5362	.6879	.9120
	weak human	.9818	.8675	.7022	.7281	.9652
	weak auto	.9845	.8557	.6634	.7150	.9479
	full	.9936	.8673	.6337	.7543	-

Table 2 Total amount of pixels annotated per class for Endovis 2015 dataset.

	full	weak auto		weak human		weak human red.	
	#	#	%	#	%	#	%
background	83,561,953	3,759,618	4.50	3,781,461	4.53	3,959,571	4.74
shaft	5,625,442	213,460	3.79	462,043	8.21	225,559	4.01
manipulator	2,972,605	174,232	5.86	323,779	10.89	171,490	5.77
total	92,160,000	4,147,310	4.72	4,567,283	7.88	4,356,620	4.84

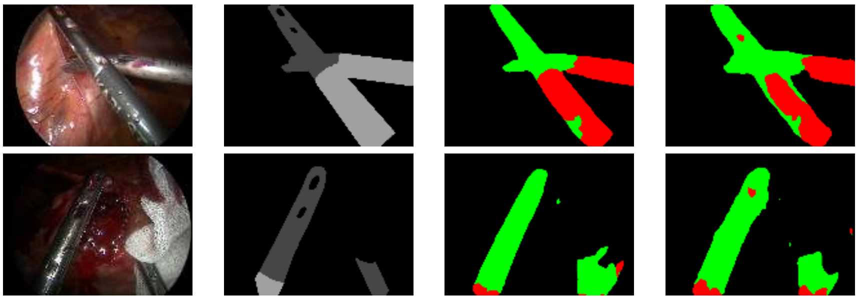

Fig. 6 Example results comparing the segmentations we obtain with our weak labels and the ones obtained with the full masks on the Endovis 2015 dataset. First column shows the original frame, second one the ground truth, third one the fully-supervised results and the last one our weakly-supervised results.

6.2 Sleeve Gastrectomy results

Table 3 compares the results obtained on the Gastric Sleeve dataset with our approach and with full supervision. In this case, with a much more complicated dataset, it can be observed that our approach gets in average 90% of the accuracy that full supervision achieves. There are 5 classes that do not have presence in every video and thus get $IOU = 0$ when these videos are in the test set. Figure 7 shows some examples of good and bad results. In addition, Table 4 shows the distribution of per-class pixel annotations for full and weak masks.

Table 3 Segmentation results for Sleeve dataset. Per-class and averaged IOU values are reported. The order of the classes is the following: *stapler, stapler handle, stapler trigger, atraumatic grasper handle, atraumatic grasper tip, liver retractor, ligasure tip, ligasure handle, marylands tip, marylands handle, bandage, liver, stomach and background*. Star (*) denotes that there exist annotated pixels for the given class. Last column shows how similar the results are to those obtained with full supervision, from 0 to 1.

fold	mask	1	2	3	4	5	6	7	8	9	10	11	12	13	14	avg	sim.
0	weak	.86	.71	.81	.82	.68	.72	.33	.75	0	0	.74	.94	.85	.92	.62	.99
	full	.95	.77	.68	.86	.75	.45	.14	.45	0	0	.81	.94	.84	.96	.63	
1	weak	.89	.89	.83	.83	.63	.32	0	0	0	0	.87	.95	.96	.90	.62	.86
	full	.97	.95	.72	.95	.60	.62	0	0	0	0	.92	.96	.94	.97	.72	
2	weak	.91	.61	.91	.77	.88	0	0	0	0*	0*	.81	.92	.95	.90	.45	.83
	full	.97	.68	.67	.86	.83	0	0	0	0*	0*	.81	.96	.96	.96	.54	
3	weak	.85	.87	.93	.84	.87	0	0	0*	0	0	.51	.52	.92	.86	.49	.80
	full	.97	.97	.56	.89	.72	0	0	0*	0	0	.67	.77	.91	.94	.61	
4	weak	.77	.68	.71	.62	.60	.15	.30	.75	0	0	.73	.76	.81	.92	.47	.96
	full	.78	.73	.52	.70	.56	.08	.15	.86	0	0	.81	.77	.76	.97	.49	
avg	weak	.86	.75	.84	.78	.73	.50	.32	.58	0*	0*	.73	.82	.90	.90	.53	.90
	full	.93	.82	.63	.85	.69	.38	.15	.44	0*	0*	.80	.88	.88	.96	.59	

Table 4 Total amount of pixels annotated per class for Sleeve dataset.

		full	weak	%
1	stapler	194,058,923	4,945,339	2.55
2	stapler handle	42,229,230	579,960	1.37
3	stapler trigger	3,723,493	390,791	10.50
4	atraumatic grasper handler	45,430,250	1,852,076	4.08
5	atraumatic grasper tip	15,273,911	1,027,326	6.73
6	liver retractor	7,566,696	333,742	4.41
7	ligasure tip	1,847,152	144,637	7.83
8	ligasure handle	3,082,359	168,734	5.47
9	marylands tip	73,861	3,732	5.05
10	marylands handle	110,951	7,513	6.77
11	bandage	126,062,766	5,004,185	3.97
12	liver	301,709,606	4,725,905	1.57
13	stomach	534,054,288	11,342,402	2.12
14	background	1,436,405,339	29,260,588	2.04
	total	2,711,628,825	59,753,930	2.20

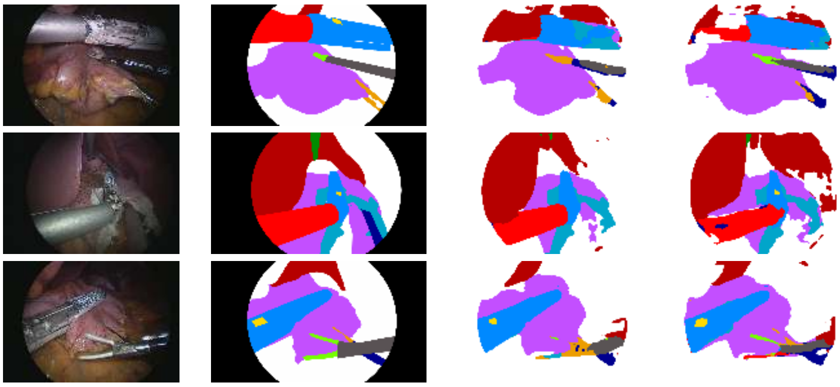

Fig. 7 Examples of good and bad results obtained for the Sleeve dataset. First column shows the original frame, second one the ground truth, third one the fully-supervised results and the last one our weakly-supervised results. As for rows, the first two rows are good results, and the last one bad results. Note how weak supervision works better than full supervision for instrument tips.

6.3 Gastric Bypass results

In this Section, we show some quantitative results for the Gastric Bypass dataset, the one which was fully annotated with our method and did not have any other ground truth available. Figure 8 shows both accurate and not so accurate results obtained by our approach for several sample images extracted from this dataset.

6.4 Time comparison

For pixel-level full annotations of the Sleeve Gastrectomy dataset, it took our team 428 (±170) seconds per image, as opposed to the 31 (±17) seconds needed with the proposed approach. Extrapolating for our complete dataset, which consists of

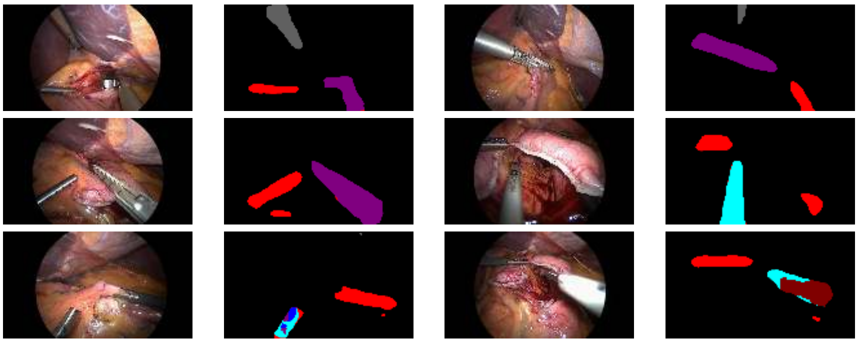

Fig. 8 Example results comparing the segmentations we obtain with our weak labels on the Gastric Bypass dataset. First two rows show good results, last one not so good.

3600 images, full annotations would need $\sim$ 53 days (assuming full dedication 8 hours/day), whereas our approach would cut that time down to $\sim$ 4 days, while keeping the performance very similar.

7 Discussion

The results obtained with full supervision for the Endovis 2015 dataset show a mean intersection over union (mIOU) of 75.43%, compared to 71.50% that we obtain with our automatically computed weak annotations, 72.81% with the manual annotations or 68.79% with the reduced manual annotations. This is very promising, as the percentage gain with full supervision is small compared to the time needed to create full labels. However, this is a rather simple dataset. Therefore, we also tested our approach with two much more complex datasets (Bypass and Sleeve). For Bypass, we obtain 59% of mIOU for full supervision and 53% for our weakly-supervised approach. These experiments supports the findings that our weak annotations only degrade the quality of the output by a small percentage difference, even for complex datasets. The per-class IOUs show that full and weak supervision obtain comparable results, and that there is a significant drop ($\sim$ 5%) when there is less data available. At the light of these results, weak segmentation appears to work better for small, surrounded shapes (e.g. liver retractor, instruments tip, stapler trigger, etc). On the other side, full segmentation generally performs better on coarse, big shapes. That might be caused by the proximity of the annotations to the boundaries: small, thin shape annotations are closer to the boundaries than those of big, coarse shapes. Although it needs further experimenting, this may prove the intuition that weak segmentation is less accurate close to boundaries, since weak masks do not hold this kind of information.

We also observe that the accuracy obtained with both datasets remains coherent, with our approach achieving between $90 - 95\%$ of the mIOU obtained by full supervision. Furthermore, we prove that our automatic annotations represent the annotations performed by a human.

The last set of experiments were performed using the Gastric Bypass dataset, for which we show qualitative results and appear to be coherent as well. In addition, if we pay attention to the first row of Figure 7, we can see as sometimes our approach is able to obtain even more accurate results than the fully-supervised approach, although this observation is anecdotal. We also observed how Endovis15 dataset works better when more annotations than just a stripe along the longitudinal axis are available, so more labeled data help to obtain more accurate results. Compared to traditional scribble-based segmentation methods, we assume our method more accurate, since these methods have been outperformed by FCNNs, and our approach achieves very similar accuracy to fully-supervised FCNNs.

8 Conclusion

The presented methodology is a step towards reducing the amount of data needed for segmentation of surgical data and speeding up the research in this field, which suffers from the lack of data. We show that we are able to obtain similar accuracy with our weakly-supervised method and with full supervision in three different datasets. However, our work has some limitations. We tested the influence of the amount of data on a very simple and small dataset, so further research must be done to find the amount of annotations needed to obtain the highest possible accuracy, and its relationship with the size of the dataset. Another drawback is that we still need fully annotated masks to test our weakly-supervised approaches. However, this is a general problem of weak supervision and is not concrete to our approach. Future lines of work include incorporating different loss functions or post-processing techniques, checking how erroneous annotations would affect our approach (compared to full supervision) and investigating if our method could be used as a previous step to full-segmentation reducing the amount of time needed. Another interesting line of work would be to try to combine our scribbles with label propagation. We could use a full mask as initialization and then compute the skeleton of it to obtain the weak mask as the labels propagate. In this way, we could allow some coarser fitting of the labels when propagating, because only the skeletons would be used when training.

Conflict of Interest. The authors declare that they have no conflict of interest.
Ethical approval. For this type of study formal consent is not required.

References

1. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: Whats the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision, pp. 549–565. Springer (2016)
2. Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., Speidel, S.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. arXiv preprint:1805.02475 (2018)
3. Bodenstedt, S., Ohnemus, A., Katic, D., Wekerle, A.L., Wagner, M., Kenngott, H., Müller-Stich, B., Dillmann, R., Speidel, S.: Real-time image-based instrument classification for laparoscopic surgery. arXiv preprint:1808.00178 (2018)
4. Bouget, D., Allan, M., Stoyanov, D., Jannin, P.: Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical image analysis **35**, 633–654 (2017)

5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Gao, M., Xu, Z., Lu, L., Wu, A., Nogues, I., Summers, R.M., Mollura, D.J.: Segmentation label propagation using deep convolutional neural networks and dense conditional random field. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), pp. 1265–1268. IEEE (2016)
8. García-Peraza-Herrera, L.C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S.: Toolnet: holistically-nested real-time segmentation of robotic surgical tools. In: Intelligent Robots and Systems (IROS), 2017 IEEE/RSJ International Conference on, pp. 5717–5722. IEEE (2017)
9. García-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Vander Poorten, E., Stoyanov, D., Vercauteren, T., Ourselin, S.: Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. In: International Workshop on Computer-Assisted and Robotic Endoscopy, pp. 84–95. Springer (2016)
10. Laina, I., Rieke, N., Rupprecht, C., Vizcaíno, J.P., Eslami, A., Tombari, F., Navab, N.: Concurrent segmentation and localization for tracking of surgical instruments. In: International conference on medical image computing and computer-assisted intervention, pp. 664–672. Springer (2017)
11. Lejeune, L., Grossrieder, J., Sznitman, R.: Iterative multi-path tracking for video and volume segmentation with sparse point supervision. Medical image analysis **50**, 65–81 (2018)
12. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3159–3167 (2016)
13. Maier-Hein, L., Ross, T., Gröhl, J., Glocker, B., Bodenstedt, S., Stock, C., Heim, E., Götz, M., Wirkert, S., Kenngott, H., Speidel, S., Maier-Hein, K.: Crowd-algorithm collaboration for large-scale endoscopic image annotation with confidence. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 616–623. Springer (2016)
14. MICCAI 2015: Endovis 2015 instrument segmentation and tracking. `https://endovissub-instrument.grand-challenge.org` (2015). [Online; accessed 6-Nov-2018]
15. Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N.: Deep residual learning for instrument segmentation in robotic surgery. arXiv preprint:1703.08580 (2017)
16. Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., Kengott, H., Speidel, S., Kop-Schneider, A., Maier-Hein, K., Maier-Hein, L.: Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery pp. 1–9 (2018)
17. Schoeffmann, K., Husslein, H., Kletz, S., Petscharnig, S., Muenzer, B., Beecks, C.: Video retrieval in laparoscopic video recordings with dynamic content descriptors. Multimedia Tools and Applications pp. 1–20 (2017)
18. Shvets, A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.: Automatic instrument segmentation in robot-assisted surgery using deep learning. arXiv preprint:1803.01207 (2018)
19. Stoyanov, D.: Surgical vision. Annals of biomedical engineering **40**(2), 332–345 (2012)
20. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City (2018)
21. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 352–368 (2018)
22. Vardazaryan, A., Mutter, D., Marescaux, J., Padoy, N.: Weakly-supervised learning for tool localization in laparoscopic videos. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis, pp. 169–179. Springer (2018)
23. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1354–1362 (2018)

24. Zhao, X., Liang, S., Wei, Y.: Pseudo mask augmented object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4061–4070 (2018)