Markus Philipp, Anna Alperovich, Alexander Lisogorov, Marielena Gutt-Will, Andrea Mathis, Stefan Saur, Andreas Raabe, Franziska Mathis-Ullrich

Annotation-efficient learning of surgical instrument activity in neurosurgery

https://doi.org/10.1515/cdbme-2022-0008

Abstract: Machine learning-based solutions rely heavily on the quality and quantity of the training data. In the medical domain, the main challenge is to acquire rich and diverse annotated datasets for training. We propose to decrease the annotation efforts and further diversify the dataset by introducing an annotation-efficient learning workflow. Instead of costly pixel-level annotation, we require only image-level labels as the remainder is covered by simulation. Thus, we obtain a large-scale dataset with realistic images and accurate ground truth annotations. We use this dataset for the instrument localization activity task together with a student-teacher approach. We demonstrate the benefits of our workflow compared to state-of-the-art methods in instrument localization that are trained only on clinical datasets, which are fully annotated by human experts.

Keywords: Annotation-efficiency learning, neurosurgery, instrument localization, medical deep learning

1 Introduction

The lack of large, annotated data is one of the main challenges in medical deep learning. This stems from the fact that the creation of such datasets is constrained by cost- and time-intensive annotations, which often require medical expertise. Annotations are especially expensive if they are on a pixel-wise level, such as segmentation or bounding boxes. To address the annotated data constraint, *annotation-efficient learning* became a relevant issue in medical deep learning [1].

We focus on the problem of localizing surgical instrument activity in neurosurgical microscope video data, see Fig. 1 (a), which is a cornerstone towards computer-assisted surgery. To train deep learning models in our prior work [2], annotators manually labelled instrument tips with bounding boxes, which we required to compute instrument activity labels, see Fig. 1 (b). Creating a medium-sized annotated dataset took hundreds of hours and many annotation rounds. To create a large-scale dataset, we need even more time and human effort. In this work, we investigate annotation-efficient learning to save annotation labour for future similar problems.

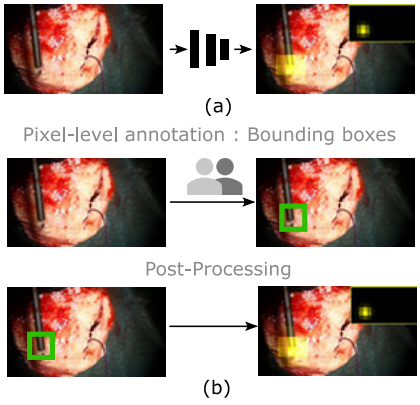


Figure 1: (a) A neurosurgical scene (left) with surgical instrument activity as yellow overlay (right). (b) Bounding box annotation for the same scene (top) and post-processing to obtain surgical activity labels (bottom).

Contributions. We propose an annotation-efficient learning workflow for surgical instrument activity localization. We abstain from costly pixel-level bounding box annotations and resort to cheaper image-level labels, which merely require annotators to decide if an instrument is present in a current frame or not. Based on these image-level annotations, we create a hybrid-synthetic data domain, where we can automatically compute instrument activity labels. In this way, we combine the advantage of human-made image-level annotations and machine-made pixel-level annotations. This approach speeds up the annotation process and diversifies the dataset with more instrument shapes and positions. Then, we formulate a student-teacher approach to learn instrument activity localization, where we use our hybrid-synthetic data domain as a proxy to guide the student. While we achieve competitive results compared to the model trained on the dataset based on costly manual bounding box annotations, our approach saves ~75% of the annotation work.

***Corresponding author: F. Mathis-Ullrich:** Health Robotics and Automation (IAR-HERA), Karlsruhe Institute of Technology (KIT), Karlsruhe, DE, e-mail: franziska.ullrich@kit.edu
M. Philipp: Health Robotics and Automation (IAR-HERA), KIT, Karlsruhe, DE & Carl Zeiss Meditec AG, Oberkochen, DE
A. Alperovich: Carl Zeiss AG, Oberkochen, DE
A. Lisogorov, S. Saur: Carl Zeiss Meditec AG, Oberkochen, DE
M. Gutt-Will, A. Mathis, A. Raabe: University Hospital Bern, CH

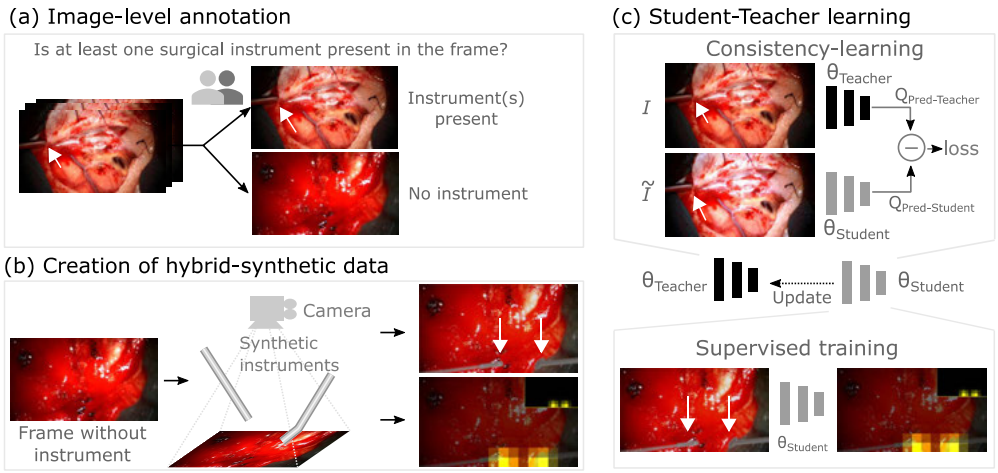


Figure 2: (a) Annotators classify if clinical video frames contain a surgical instrument or not (image-level annotation). For better visibility, we mark instrument tips with a white arrow. (b) Frames without instruments are used to create hybrid-synthetic data. We overlay synthetic instruments as foreground and compute the according label Q*. (c) The student-teacher learning incorporates annotated hybrid-synthetic data and unlabeled clinical data.

1.1 Related work

Current approaches to surgical instrument localization address annotation efficiency in different ways: [3] boost instrument segmentation through a self-supervised pre-training on unlabelled surgical data. [4] follow a different approach and apply weak supervision to simplify the annotation labour from segmentation level to stripe level. Image-to-image techniques are leveraged in [5] for style transfer between labelled and unlabelled datasets. On the other hand, [6] use domain adaption to combine rendered, synthetic laparoscopic data [7] with unlabelled clinical data. However, applying such an approach to the neurosurgical domain is currently impossible since no synthetic dataset as [7] is available.

We build upon [6] and introduce a hybrid-synthetic data domain. We refer to hybrid-synthetic data as a mixture of real-world clinical background images and synthetic instruments overlaid as foreground. Hybrid-synthetic data tackles various challenges of *purely* synthetic data: (1) no complex surgical scene/anatomy modelling is required, (2) high variability can be achieved easily by exchanging the background, (3) realistic appearance due to real-world clinical backgrounds; thus, smaller domain gap to the real-world clinical test domain.

2 Method

We consider the problem of predicting surgical instrument activity as a 16 x 9 saliency map $Q = (p_{i,j})$, where $p_{i,j}$ describes the probability for an instrument tip in the image region (i,j), see Fig. 1 (a). Our goal is to train a model θ that can infer instrument activity Q for an image input I. Thereby, I comes from a real-world clinical domain, $\mathcal{D}_{clinical}$.

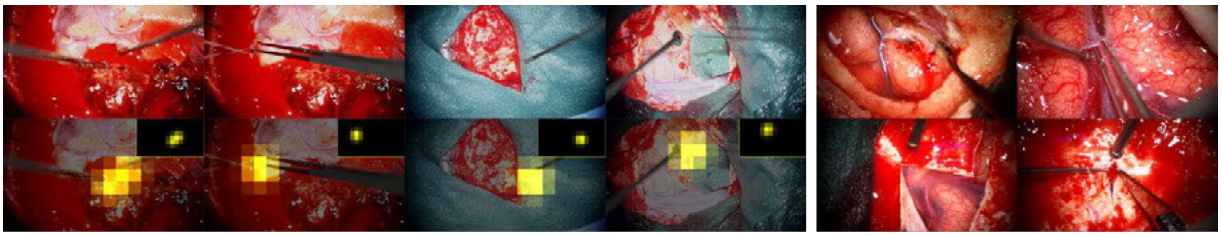
To train θ in a supervised fashion as in [1], one needs clinical training data $\{I, Q^*\}_{clinical}$ with $I \in \mathcal{D}_{clinical}$, that consist of images I with corresponding reference labels Q^*. To create Q^*, bounding box annotations are needed (Fig. 1).

Our method avoids the need for manually labelling bounding boxes. Instead, we use cheaper image-level annotations created based on the question of whether annotators see surgical instruments in the frame. We employ these image-level annotations to design a hybrid-synthetic domain $\mathcal{D}_{hybrid}$ which we define such that we can automatically compute Q^*. This allows to leverage the benefits of human-made manual image-level annotations and machine-made pixel-level annotations. Finally, we take labelled data from $\mathcal{D}_{hybrid}$ and unlabelled data from $\mathcal{D}_{clinical}$ to learn instrument localization based on a student-teacher approach.

In summarizing, our method consists of three steps: (a) Conduct image-level annotations, (b) based on them, create hybrid-synthetic data, (c) train a model θ using a student-teacher approach. We give an overview of our method in Fig. 2 and describe its steps in more detail in the following sections.

2.1 Image-level annotations

For our image-level annotations, annotators classify if the surgical instruments are present in a frame or not. Our observations show that such image-level annotations take only approx. 25% of the time required for bounding box annotations. We assume to have such image-level annotations for a set of real-world clinical images $\{I\}_{clinical} \subseteq \mathcal{D}_{clinical}$. Based on this presence/absence annotation, we divide $\{I\}_{clinical}$ into two subsets, $\{I\}_{presence}$ and $\{I\}_{absence}$.



(a) (b)

Fig 3: (a) We show different samples of our hybrid-synthetic dataset. Upper row: generated images, lower row: labels $\boldsymbol{Q}^*$. (b) We show different samples from the clinical data. Comparing (a) and (b) confirms the realism of our hybrid-synthetic data.

2.2 Hybrid-synthetic neurosurgical data

Our goal is to synthesize neurosurgical training data, where we can obtain saliency labels Q^* at no additional annotation cost. We achieve this by creating a hybrid-synthetic data domain $\mathcal{D}_{hybrid}$. Based on image-level labels, we design $\mathcal{D}_{hybrid}$ such that we can automatically compute saliency labels Q^*.

We create hybrid-synthetic neurosurgical data by using our framework presented in [8]. It uses the 3D animation software Blender to render neurosurgical scenes in two steps: (1) Generate a random geometric constellation of neurosurgical instruments, (2) underly single neurosurgical microscope images as a background. Fig. 2 (b) illustrates the creation of such hybrid-synthetic data. For details, see [8].

We employ the image-level annotation to design $\mathcal{D}_{hybrid}$ such that we can automatically compute Q^*: As background images, we take images from $\{I\}_{absence}$, as they do not contain surgical instruments already. When adding synthetic instruments on-top of $\{I\}_{absence}$ images, the synthetic instruments are the only instruments in the rendered image. Consequently, we can automatically compute labels Q^* from the position of the synthetic instruments in Blender. We render a dataset $\{I, Q^*\}_{hybrid}$ and show samples from it in Fig. 3 (a). We provide reference of *real* clinical images in Fig. 3 (b).

2.3 Student-Teacher-Learning

Our goal is to train a model θ for the clinical domain $\mathcal{D}_{clinical}$. Building upon [6], we formulate a student-teacher task for the regression problem of instrument activity localization. Our approach combines supervised training on the labelled hybrid-synthetic domain $\mathcal{D}_{hybrid}$ and domain adaption to the unlabelled clinical domain $\mathcal{D}_{clinical}$ through consistency learning. By this way, two networks - a student network $\theta_{student}$ and a teacher network $\theta_{teacher}$ with identical architecture - interact in a two-step cycle:

Step 1. The teacher network receives an input image I sampled from $\{I\}_{clinical}$ to predict $Q_{pred-teacher}$. As the true label Q^* is unknown, $Q_{pred-teacher}$ serves as a *pseudo-label* for the student. The student is trained on two tasks simultaneously:

(a) Consistency learning: The goal of the consistency loss is to make student familiar with variations in the clinical domain $\mathcal{D}_{clinical}$ which are simulated by data augmentations. The same image I is pixel-wise perturbated (see Fig. 2 (c)) to obtain $\tilde{I}$. $\tilde{I}$ is given as input to the student. Since pixel-wise perturbations have no effect on Q^*, the student's output $Q_{pred-student}$ should match $Q_{pred-teacher}$. This is enforced by a consistency loss:

$$\mathcal{L}_{consistency} = \left(Q_{pred-student} - Q_{pred-teacher}\right)^2 \quad (1)$$

(b) Supervised task learning: To focus the student on the instrument activity localization task, the student is trained on labelled hybrid-synthetic data $\{I, Q^*\}_{hybrid}$ with:

$$\mathcal{L}_{supervised} = \left(Q_{pred-student} - Q^*\right)^2, \quad (2)$$

whereas $Q_{pred-stud}$ is the student's prediction. The student's weights θ_{stud} are updated based on combined loss as in [6],

$$\mathcal{L}_{student} = \alpha(t) * \mathcal{L}_{consistency} + \mathcal{L}_{supervised}, \quad (3)$$

with a weighting factor $\alpha(t)$. The weighting factor is increased throughout the training to shift the student's focus from $\mathcal{D}_{hybrid}$ to $\mathcal{D}_{clinical}$. We use the same loss function for the supervised training as for the consistency training to allow a smooth transition between the task with increasing $\alpha(t)$.

Step 2. We update the teacher by an exponential filter as [6]:

$$\theta_{teacher} = 0.95 * \theta_{teacher} + 0.05 * \theta_{stud} \quad (4)$$

By repeating this two-step cycle, the student benefits from an improved teacher due to better pseudo-labels.

3 Experiments

We describe our experimental setup. Then, we explain baselines that serve as a comparison to our method.

Data. We use the *NeuroSurg* dataset introduced in [2], for which both instrument presence/absence labels and instrument activity labels are available. We ignore the instrument activity labels for training our method and only rely on the instrument presence/absence labels. We consider six neurosurgical cases for training purposes and test on the six independent cases [2].

Evaluation. We use the SIM metric, which is a standard metric in the saliency literature: $SIM = \sum \min\left(Q_{pred}, Q^*\right)$, whereas $\sum Q_{pred} = \sum Q^* = 1$. *Hybrid-synthetic data.* Based on the image-level annotations for the six training surgeries, we generate a dataset $\{I, Q\}_{hybrid}$ with 20.000 training images.

Table 1: We test our method and the baselines on the six test cases (Case No. 1 – 6) from NeuroSurg. We report mean SIM values and standard deviations for the six test cases and mark the best mean SIM values per case bold.

	Case No. 1	Case No. 2	Case No. 3	Case No. 4	Case No. 5	Case No. 6
Our method	0.76±0.14	0.75±0.17	0.75±0.16	0.67±0.20	0.77±0.12	0.67±0.18
Clinical6-FS	**0.83±0.11**	**0.81±0.15**	**0.78±0.12**	**0.72±0.18**	**0.78±0.12**	**0.72±0.13**
Clinical2-FS	0.72±0.17	0.67±0.19	0.68±0.16	0.63±0.20	0.71±0.15	0.68±0.16
Hybrid-FS	0.67±0.22	0.69±0.25	0.71±0.21	0.59±0.26	0.73±0.17	0.58±0.24

Student-Teacher-Learning implementation. We use the CNN architecture from [2] and re-implement the perturbations as in [6]. Initial learning rate lr=0.01 (reduced by 0.5 every 50 epochs), no. of epochs = 300, batch size = 25, $\alpha(t) = \{0$ for $t \leq 10$, lin. increase to 1 for $t \in [11,50]$, 1 for $t \geq 50\}$.

Baseline. We investigate the benefit from annotating the training data with bounding boxes as in Fig. 1. Also, we explore the advantages of the student-teacher approach in contrast to mere supervised training on the hybrid-synthetic data $\{I, Q\}_{hybrid}$. We compare our method with several baselines: (1) We use fully supervised model from [2] which we refer to as *Clinical6-FS*. (2) To investigate the effect of training dataset size, we train a supervised baseline, *Clinical2-FS,* on only two training cases of *NeuroSurg* (~31% of the training data of *Clinical6-FS*). (3) We train another baseline *Hybrid-FS* on $\{I, Q\}_{hybrid}$ using the conditions as in [2].

4 Results

We compare the performance of our annotation-efficient learning method with the two baseline methods in Tab. 1.

Our annotation-efficient learning method achieves a competitive performance to the baseline *Clinical6-FS.* It performs close to on-par on some test cases and slightly worse on the remaining cases. Our method even outperforms the supervised baseline *Clinical2-FS* on five of six test cases.

Now we compare *Hybrid-FS* and *Clinical6-FS.* Although *Hybrid-FS* never saw *real* instruments – only real clinical backgrounds - during training, it continuously achieves > 80% of performance of *Clinical6-FS*. This supports our claim that our hybrid-synthetic data are highly realistic.

Finally, we compare our method with *Hybrid-FS.* Our method outperforms *Hybrid-FS* on all test cases. Despite the already good performance of *Hybrid-FS* on the test data, we still gain benefits from the student-teacher approach.

5 Conclusions

We leverage hybrid-synthetic data and a student-teacher learning approach for annotation-efficient learning of surgical instrument activity. Our approach replaces effort- and cost-intensive bounding box annotations with simpler and cheaper image-level annotations. We demonstrate how to generate a realistically looking large-scale synthetic dataset for training by successfully combining human-made and machine-made annotations. While we achieve a competitive performance compared to state-of-the-art supervised learning based on bounding box annotations, we save up to 75% annotation effort.

Author Statement

Research funding: The author state no funding is involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration.

References

[1] Tajbakhsh, N. et al. (2021). Guest Editorial Annotation-Efficient Deep Learning: The Holy Grail of Medical Imaging. IEEE T-MI, 40(10), 2526-2533.

[2] Philipp, M. et al. (2021). Localizing Neurosurgical Instruments Across Domains and in the Wild. MIDL 2021.

[3] Ross, T. et al. (2018). Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int J CARS, 13, 925-933.

[4] Fuentes-Hurtado, F. et al. (2019). EasyLabels: weak labels for scene segmentation in laparoscopic videos. Int J CARS, 14, 1247-1257.

[5] Kalia, M. et al. (2021). Co-Generation and Segmentation for Generalized Surgical Instrument Segmentation on Unlabelled Data. MICCAI 2021.

[6] Sahu, M. et al. (2021). Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation. Int J CARS 16, 849–859.

[7] Pfeiffer, M. et al. (2019). Generating Large Labeled Data Sets for Laparoscopic Image Processing Tasks Using Unpaired Image-to-Image Translation. MICCAI 2019.

[8] Philipp M., et al. (2021). Synthetic data generation for optical flow evaluation in the neurosurgical domain. Curr. Dir. Biomed. Eng, 7(1), 67-71