# Leveraging active learning techniques for surgical instrument recognition and localization

**Bartłomiej Piotrowski**[1],* **Jakub Oszczak**[1], **Krzysztof Sawicki**[1], **Barbara Siemiątkowska**[1] ,***Andrea Curatolo**[2,3] ***

[1] Institute of Automatic Control and Robotics, Warsaw University of Technology, A. Boboli 8, 02-525 Warsaw, Poland;
[2] International Centre for Translational Eye Research, Skierniewicka 10A, 01-230 Warsaw, Poland;
[3]Institute of Physical Chemistry, Polish Academy of Sciences, Kasprzaka 44/52, 01-224 Warsaw, Poland;

**Abstract.** The field of ophthalmic surgery demands accurate identification of specialized surgical instruments. Manual recognition can be time-consuming and prone to errors. In recent years neural networks have emerged as promising techniques for automating the classification process. However, the deployment of these advanced algorithms requires the collection of large amounts of data and a painstaking process of tagging selected elements.

This paper presents a novel investigation into the application of neural networks for the detection and classification of surgical instruments in ophthalmic surgery. The main focus of the research is the application of active learning techniques, in which the model is trained by selecting the most informative instances to expand the training set. Various active learning methods are compared, with a focus on their effectiveness in reducing the need for significant data annotation - a major concern in the field of surgery. The use of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to achieve high performance in the task of surgical tool detection is outlined. The combination of Artificial Intelligence (AI), Machine Learning, and Active Learning approaches, specifically in the field of ophthalmic surgery, opens new perspectives for improved diagnosis and surgical planning, ultimately leading to an improvement in patient safety and treatment outcomes.

The article consists of seven sections. After the introduction, the classic approach to surgical instrument recognition is presented. In sections 3-4 the neural networks and active learning techniques are described. Our algorithm for surgical tool recognition is presented in sec. 5. The article concludes with a discussion.

**Key words:** surgical tool detection, eye surgery, neural networks, data annotation

## 1. INTRODUCTION

Today, technological advances have a profound impact on many aspects of life and science, including healthcare. Eye surgery, one of the most precise and demanding areas of medicine, is benefiting from technological innovations to improve the efficiency of procedures and minimize risks to patients.

In the field of ophthalmic microsurgery, we are witnessing a growing interest and implementation of new technologies that are expected to increase the effectiveness of surgical procedures and minimize the risk of postoperative complications as well as errors during the operation itself. In this context, scientific fields such as computer vision and machine learning are becoming more important and are increasingly being used in ophthalmic surgery.

During ophthalmic procedures, surgeons face a variety of challenges and potential risks. An important issue to consider is the risk of intraocular hemorrhage, which can significantly complicate the precise positioning of surgical instruments. Subretinal surgery demands particular attention, as inaccurate instrument positioning can result in damage to the optic nerve, with fundamental implications for the patient's perceptual abilities.

Retinal microsurgery requires extremely high levels of visual acuity and depth perception, as well as precise hand movements. The required precision for positioning the tool in the eye is 10 $\mu$m under general conditions, and 25 $\mu$m during sub-retinal injection [1, 2].

A person with "normal" vision, defined by the Snellen test as 20/20, has the ability to distinguish two points in his field of vision as separate if they are separated by an angle of at least 1 minute of arc. In practice, this means that at a distance of 1 meter, the eye can distinguish two points that are separated by at least 0.29 mm. At a distance of 20 meters, this resolution is approximately 5.8 mm [3].

Microscopic magnification can significantly improve visual resolution. However, despite using a microscope, the precision of human visual perception cannot match the accuracy of modern cameras - especially those integrated with the microscope sharing the same optical path as the surgeon.

The resolution capability of optical systems in microscopes greatly surpasses the visual capability of the human eye. This offers immense potential for improving the precision of tracking surgical instruments and other complex tasks. The application of such sophisticated optical technology can significantly improve the accuracy and efficiency of medical procedures, a critical aspect in optimizing patient outcomes.

This article focuses specifically on advancing the application of deep learning techniques, in particular active learning, to improve the accuracy of surgical instrument localisation and segmentation in ophthalmic surgery. Our work builds on initial developments using the OpenCV library, which have been instrumental in real-time tracking of the tip of surgical instruments during procedures. While these traditional methods pro-

---

*e-mail: bartlomiej.piotrowski.dokt@pw.edu.pl
**e-mail: barbara.siemiatkowska@pw.edu.pl
***e-mail: acuratolo@ichf.edu.pl.

Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska

vided valuable insights, they were limited in their scope and adaptability to different surgical environments.

The initial program developed using OpenCV focused primarily on real-time localisation and tracking of the tip and orientation of the surgical tool during eye surgery. This early approach laid the foundation for understanding the dynamics of surgical instruments in a controlled environment. **However, the method can be ineffective because of its limited ability to adapt to the varying complexity of different surgical scenarios.**

In response to these limitations, our current research explores the use of neural networks to go beyond the capabilities of the OpenCV-based approach. We are focusing on using neural networks to not only track, but also accurately localise and segment surgical instruments within different surgical environments. By integrating active learning into this approach, we aim to create a more flexible and robust system that can adapt to different types of eye surgery and conditions, thereby improving the overall precision and effectiveness of surgical procedures.

The article is structured into the following sections:

1. **Introduction:** Discusses the impact of technological advances on health care, particularly in the field of ophthalmic surgery. It highlights the challenges and potential risks surgeons face during ophthalmic procedures and the need for precise instrument positioning.
2. **From Traditional Methods to Active Learning in Ophthalmic Instrument Recognition and Localisation:** This section outlines the limitations of traditional methods for surgical instrument recognition, setting the stage for the introduction of advanced techniques such as CNNs and active learning.
3. **Convolutional Neural Networks:** This section explains how CNNs have revolutionized image processing by enabling accurate and efficient classification, but also highlights the challenge of the need for a substantial amount of labeled training data.
4. **Active Learning Methods:** Active learning, a type of supervised learning that involves the model in selecting the most informative samples, is covered in this section. It is introduced as an effective way to reduce the need for extensive data annotation.
5. **Algorithms for Surgical Instrument Recognition:** This section describes the YOLOv5-based algorithm, enhanced with active and semi-supervised learning, for surgical instrument detection, using a dataset annotated via Roboflow.
6. **Experimental Results:** This section discusses four different experiments that use active learning and semi-supervised learning methods to reduce manual image labeling. It provides a detailed analysis of methods including self-training, active learning with multiple confidence thresholds, and progressive reduction of automatic labeling thresholds.
7. **Conclusions and Future Research:** This section highlights the effectiveness and potential limitations of active learning in surgical image recognition. It also considers how this approach can be improved by integrating semi-supervised

learning and advanced tool localization methods.

## 2. ADVANCED METHOD IN OPHTHALMIC INSTRUMENT RECOGNITION AND LOCALISATION

Initially, in our attempts to develop advanced methods for surgical instrument recognition in ophthalmic surgery, we started by using traditional image processing techniques, utilising the capabilities of the OpenCV library [4]. The purpose of this **section** is to present these initial methods as a foundation for the more advanced techniques that we have subsequently developed. By analysing the capabilities and limitations of these conventional approaches, we highlight the need to move towards machine learning-based solutions in the variable and dynamic conditions of surgical environments.

In our initial efforts, we focused on manipulating colour spaces in images, using formats such as Lab* and HSV, which allowed for improved differentiation and segmentation of objects. These colour spaces have been widely used in object segmentation tasks, as evidenced by studies such as [5].

The starting point for these initial efforts was the development of a classical surgical instrument segmentation algorithm [6]. **Based on our project, our research team has developed and implemented a method that utilizes a combination of masks from both color spaces. This method is able to effectively isolate the surgical tool from the background of the image. Additionally, we deliberately applied an edge detection technique that is insensitive to shadows cast by the surgical tool, which, combined with the mask method from both color spaces, increases the accuracy of our algorithm.**

**However, these classical algorithms, while effective under controlled conditions, lacked the flexibility required for diverse operating conditions.** This limitation, evidenced in the works of Lin et al. [7], Luijten et al. [8], was apparent in their inability to generalize across various surgical scenarios and instruments. In particular, it could only be precisely parameterised to recognise the position of a single surgical tool under the specific conditions of a single operation. The inability to easily generalise the algorithm's settings to different surgeries, different lighting conditions, and different surgical tools highlighted its limitations. This inflexibility underscored the need for a more adaptive and versatile approach, which led to our exploration of machine learning-based methods.

Our research also considered alternative traditional methods for instrument tracking and localization. Studies such as those by Allan et al. [9], Bouget et al. [10, 11], Zhou [12], Sznitman et al. [13], and Rieke et al. [14, 15] provided insights into various non-ML techniques like feature-based tracking and pattern recognition, which have been instrumental in advancing surgical tool detection and tracking.

In summary, while the classical algorithm provided a solid foundation and demonstrated potential in specific scenarios, particularly in the case of surgical tool localisation and orientation, its inability to generalise across different surgical conditions and instruments necessitated the transition to more robust, adaptive solutions offered by machine learning techniques. This transition marks an important step in addressing

2

the complexity and variability inherent in live surgical environments, paving the way for more accurate and versatile surgical instrument recognition.

## 3. CONVOLUTIONAL NEURAL NETWORKS

In recent years, neural networks, particularly Convolutional Neural Networks (CNNs), have revolutionized image processing, enabling accurate and efficient classification [16]. CNNs consist of three fundamental types of layers: convolutional layers, pooling layers, and fully connected layers. The convolutional layers apply filters to detect various features, such as edges or textures. Pooling layers downsample the data, reducing its dimensionality while retaining the essential information. Finally, fully connected layers use the extracted features to predict and classify the image. Automatic feature extraction is one of the main advantages of CNNs. Unlike classical methods, which rely on predefined features, neural networks learn to identify key elements directly from raw input data. CNNs learn to extract increasingly complex and high-level features and to find relationships within the image data. This process leads to improved classification accuracy. YOLO [17] (You Only Look Once) is an example of CNN. The algorithm has gained significant popularity in the field of computer vision. One of the main advantages of YOLOv5 [18] is its speed and efficiency in real-time object detection. YOLOv5 achieves impressive performance on modern hardware, enabling it to be deployed in applications where low latency is crucial.

Despite these advantages, neural networks also have disadvantages - CNNs require a substantial amount of labeled training data to achieve optimal performance. Acquiring and annotating massive datasets can be time-consuming and challenging.

## 4. ACTIVE LEARNING METHODS

Active learning is a specific type of supervised learning, which implies the control of a person over the process of teaching [19, 20, 21]. Instead of passively relying on a fixed, labeled dataset, active learning actively involves a model in the process of selecting the data.

Initially, the neural network model is trained on a small labelled dataset, typically consisting of randomly selected or expert-labelled examples. Then, in the active learning phase, our algorithm selects unlabelled instances from a larger dataset **where the classifier exhibits high uncertainty regarding their classification.** These newly labelled instances are then added to the training dataset, and the model is retrained to improve its accuracy and generalisation ability. This iterative process of selecting informative samples **based on classifier uncertainty** and retraining the model is the essence of active learning.

In the following step, the active learning algorithm selects a subset of instances from the unlabeled dataset based on their estimated uncertainty. The goal is to choose the most informative samples that can improve the model's performance. The samples can be chosen using metrics such as entropy or margin sampling. The query strategies are described in section 4A.

Figures 1a-1c present the idea behind the method. Fig. 1a shows an example dataset divided into two classes. Fig. 1b presents the initial, randomly selected set of examples used in the learning process and the generated discrimination function. In Fig. 1c, the red highlighted examples are those selected close to the decision boundary due to their high informative value for retraining the model. Not all examples near the decision boundary are highlighted because the selection is strategic, not random, focusing on instances where the model's uncertainty is highest. These selected samples are then labeled by human annotators and included in the training dataset to improve the model's accuracy. This selective process is a critical aspect of active learning, aimed at efficiently using data to refine the model.

The **stopping criteria** of the algorithm depends on the application. Usually, one of the following methods is used:

- Budget-based criteria - a limit on the number of samples or annotations is defined.
- Model performance criteria - the process terminates when the model reaches a satisfactory performance level or performance improvement falls below a certain threshold.
- Convergence criteria - the process stops when the performance stabilizes or when there is no significant improvement in the model's performance after a certain number of iterations.

Sometimes, stop criteria can be customized based on specific requirements or domain knowledge.

The key premise of active learning is the ability to train a classifier, such as a neural network, using less data and without the need for tedious, hours-long labeling.

In addition, in the context of our research on surgical tool recognition, active learning is used not only to make efficient use of a limited data set, but also to improve the model under more diverse surgical conditions. As the previously developed OpenCV-based model was limited in scope, our active learning approach aims to extend the model's ability to work in different surgical scenarios and with different surgical tools.

### A. Query strategies

The common goal of all query strategies is to select the most relevant, information-rich data that will provide the model with the fastest growth.

Typically, one of the approaches is employed:

- Uncertainty Sampling - selects instances based on the model's uncertainty in its predictions.
- Least Confidence - the instances where it is least confident about its predictions are chosen. **It selects data points where the predicted probability of the chosen class is lowest.**
- Margin Sampling - the model focuses on instances close to the decision boundary, which are likely more informative. **Margin sampling is similar to least confidence sampling, but it considers the probability difference between the top two predicted classes. It selects instances with the**
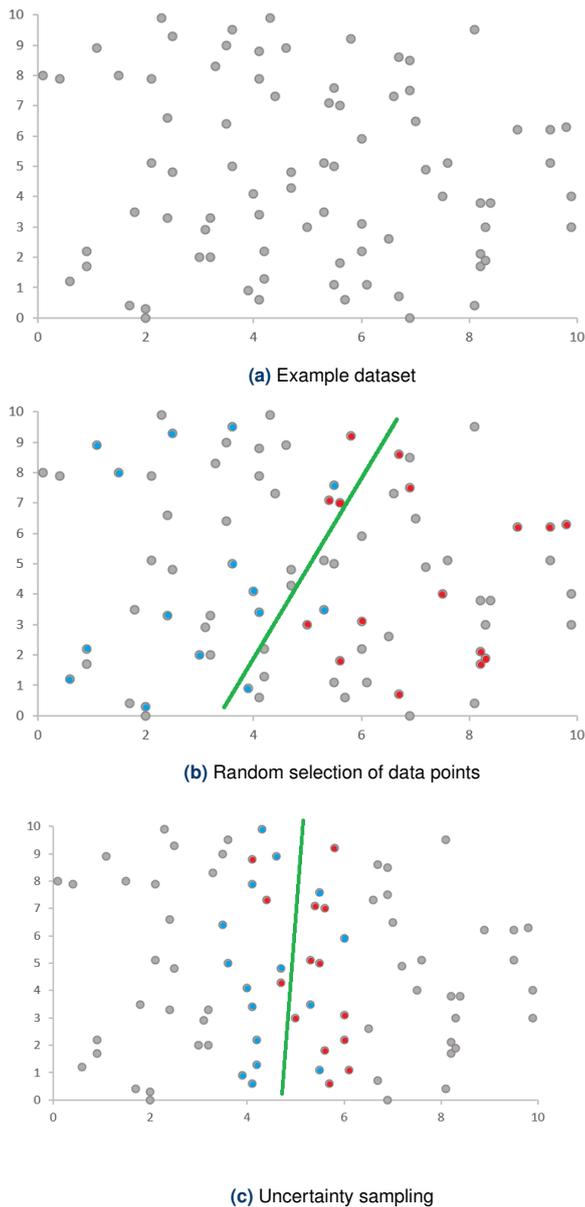
Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska



**(a)** Example dataset



**(b)** Random selection of data points



**(c)** Uncertainty sampling

**Fig. 1.** Figure represents the results of selecting data points in a random fashion and using uncertainty sampling strategy [19]

**smallest margin between the top two predicted probabilities.**

- Entropy - the entropy of the predicted class probabilities is calculated, and instances with high entropy (i.e., higher uncertainty) are selected.
- Query by Committee - this strategy involves training an ensemble of multiple models on the labeled dataset and selecting instances where the models disagree or are uncertain. The disagreement can be measured using various techniques, such as vote entropy or KL-divergence [22].
- Expected Model Change - the method selects instances that are expected to cause the most significant change in the model's parameters or predictions when added to the labeled dataset. This can be measured by computing the gradient of the model's parameters with respect to the instance's input

or using heuristic approaches like uncertainty change or loss change.

Besides query strategies, there are also query scenarios that dictate the way the model receives the data. There are three common query scenarios [19]:

- Query synthesis - in this case, the algorithm can simply send a query for an existing instance in the database, but it also has the ability to create a new data sample based on the available data and on the known constraints. It is most desirable to create samples that lie around the decision boundary of the algorithm. Query synthesis can be useful in some cases, but generally cannot be used effectively, especially in issues related to images [23]. This is because generating new images that make sense and are interpretable by humans is an extremely difficult task, for which the generative adversarial networks (GANs) are used. It is useful to refer to Lang and Baum's work [24], in which this approach was used to teach a neural network to recognize handwritten digits in order to visualize the problem.
- Pool-based sampling - this approach assumes the presence of a large dataset containing mostly unlabeled data and a small portion of labeled data. Due to the fact that it is the most frequent data format when working with machine learning, it is the most common type of query scenario. The data is sent to the algorithm in the form of a batch or all at once.
- Sequence-based selective sampling - as the name suggests, this scenario implies that a sequence of input data is available. The algorithm takes data samples from the sequence one at a time, one by one, and for each one, it decides whether to accept the sample for labeling or reject it. The decision-making process is based on query strategies. This approach assumes that the acquisition of unlabeled data has no cost.

The strategy used in this research paper is a variant of the uncertainty sampling strategy called Least confidence. In the case of unbalanced data, the Least Confidence strategy can help address the issue by actively selecting samples from the minority class that the model is least confident about.

## 5. ALGORITHMS FOR SURGICAL INSTRUMENT RECOGNITION

strument recognition algorithm we developed is an active learning method applied to the Yolov5 neural network. The algorithm is further enhanced with a semi-supervised learning method. In the following subsections, the steps of our proposed method will be described in detail.

### A. Dataset Collection

Our dataset was collected at the ICTER – International Centre for Translational Eye Research, where we recorded images of several laboratory and surgical instruments. The data belongs to five classes. The names of the instruments are presented in tab. 1. Fig. 2 is the photo of instruments used in our system.

4

**Table 1.** Class balance

| class | number of photos |
|---|---|
| CURVED_CANNULA | 750 |
| ILM_FORCEPS | 675 |
| DALK_CORNEAL_DISSECTOR | 650 |
| TITANIUM_FORCEPS | 647 |
| PEELING_SPATULA | 600 |
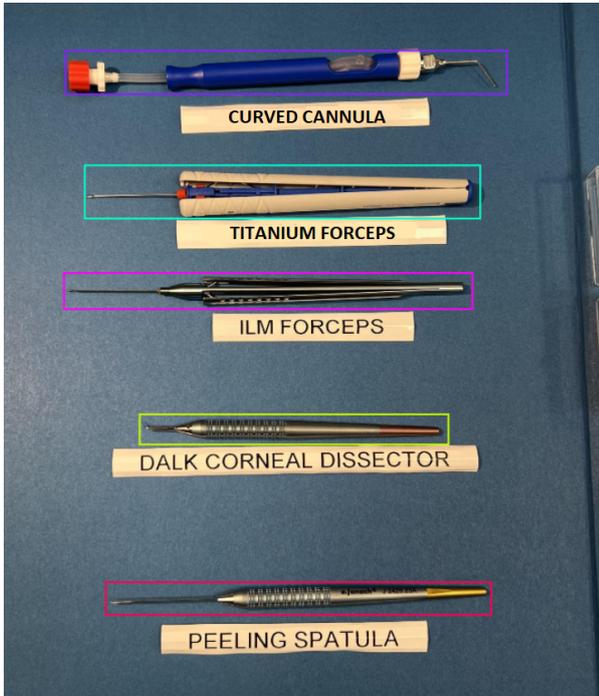


**Fig. 2.** Surgical instruments used in our system. *ILM: Retinal Internal Limiting Membrane

Data was collected with a mirrorless camera - Sony A7III and an iPhone 13 pro. In order to vary the data as much as possible, lighting conditions and backgrounds were changed during the shooting. Photos needed to be reformatted and compressed. In total, there were 3476 images, including frames extracted from the recorded videos, of which 211 had no surgical instrument, only the background. The table below shows the distribution of the number of photos per class.

### B. Dataset Annotation

The photos were labeled using the Roboflow tool. We chose the Bounding Box method because of the ability to accurately contain the entire tool in the bounding box, the speed and ease of labeling, and the fact that YOLO detection algorithm supports such a format. Other methods, such as polygonal annotation and semantic segmentation, could fail to capture a small piece of the tool which would have a harmful effect on the accuracy of the predicted bounding boxes.

### C. YOLOv5 algorithms for Surgical Instrument Recognition

We have decided to use YOLOv5 (You Only Look Once) object detection algorithm because of its popularity and excellent performance. Its full name derives from the innovative way the algorithm works. Detection and classification of objects are done with a single image pass through the network. Such single-shot algorithms have a bit worse quality of detection, but they work faster than double-shot detectors. The network consists of the following layers:

- Backbone - refers to the initial part of the network responsible for feature extraction from the input image. This task is done using CSPDarknet53 (CSP stands for Cross-Stage Partial) artificial neural network pre-trained on the ImageNet dataset that extracts relevant features from an image. It is used to reduce the spatial resolution and to increase the resolution of image features (higher number of channels).
- Neck - includes the SPP (Spatial Pyramid Pooling) and PANet (Path Aggregation Network) layers. Thanks to the SPP layer, the artificial neural network can operate effectively on input data of different dimensions [25]. PANet has been used because of its ability to preserve spatial information, making it possible to accurately locate pixels belonging to a particular class. PANet can ensure the efficient flow of spatial information from the lower layers to the final layers of the network.
- Detection Head - is responsible for predicting bounding boxes and class probabilities for detected objects. Each detection layer is associated with a specific scale of the feature map from the neck.
- Predictions - consist of bounding boxes for detected objects along with their corresponding class labels and confidence scores. Non-maximum suppression (NMS) is applied to remove duplicate and low-confidence detections.

YOLO uses Swish activation functions in the hidden layers and a sigmoidal function in the output layer.

Since YOLO returns three output values - class prediction, bounding box, and confidence score, the total loss can be expressed by the formula:

$$loss = \lambda_1 L_{cls} + \lambda_2 L_{loc} + \lambda_3 L_{obj} \qquad (1)$$

where: $\lambda_i$ - the weight of a particular type of loss, $i = 1, 2, 3$
$L_{cls}$- class prediction loss,
$L_{loc}$ - bounding box loss,
$L_{obj}$ - prediction confidence loss.

For the class prediction and confidence score, Binary Cross-Entropy was used as the loss function. For the prediction of bounding box location, Complete Intersection over Union (CIoU) was used as the loss function [26].

Yolo offers several sizes of its yolov5 algorithm. The biggest one - yolov5x has the most parameters. It is the most computationally intensive, takes the longest to train and detect, but has the best performance.

We have tested mAP metric and the average time of training and detection for each yolov5 model. We did not use active learning at this stage. While mAP (mean Average Precision) is

5

Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska

**Table 2.** Yolov5 models performance comparison

| YOLOv5 model | Trainig time (avg) | Detection time (avg) |
| --- | --- | --- |
| YOLOv5s | 1.07h | 9.9ms |
| YOLOv5m | 2.05h | 17.1ms |
| YOLOv5l | 3.67h | 30.5ms |
| YOLOv5x | 5.10h | 41.2ms |



**Fig. 3.** mAP-0.5:0.95 metric for yolov5 models

typically associated with segmentation models, in the context of this study, we use it to evaluate the object detection performance of YOLOv5. This is because mAP effectively measures the model's accuracy in localizing and classifying surgical instruments across various IoU (Intersection over Union) thresholds. It's a comprehensive metric that captures both the precision and recall of the model, making it suitable for assessing the performance of YOLOv5 in our specific case of surgical instrument detection.

The mAP metric is based on PR-curve (curve that shows precision and recall for different confidence thresholds) and it is an AP (weighted average of precision for different confidence thresholds, with weight being the change in recall between those thresholds) averaged over different IoU (intersection over union) thresholds and all classes. Below is the formula for mAP metric:

$$mAP = \frac{1}{N}\sum_N \left( \frac{1}{M}\sum_M \left( \frac{1}{\sum_i w_i}\sum_i P_i w_i \right) \right) \quad (2)$$

where:

$N$ − number of classes,

$M$ − number of IoU thresholds,

$P_i$ − precision for the i-th prediction confidence threshold,

$w_i$ − weight for the i-th prediction confidence threshold

(weight being the change in recall relative to the

previous threshold (i-1)).

The mAP_0.5:0.95 takes into account the average AP for IoU values from 0.5 - 0.95 with a step of 0.05.

Training set consisted of 80% of our dataset. Obtained results are shown in table 2.

The figure 3 is a comparison of mAP metrics for different YOLOv5 models throughout the learning process and it clearly shows the dominance of YOLOv5 m, l and x over the s model.

Finally, we decided to choose the largest yolov5x model. The decision was made upon the fact that in medical situations, detection performance is absolutely crucial, and yolov5x delivered that. It could also run in 24 frames per second which is sufficient it real-time detection.

## 6. EXPERIMENTAL RESULTS

Our objective was to test active learning methods supported by the model self-learning technique to reduce the number of images that have to be labeled manually. Additionally, the self-training algorithm has been implemented and tested. Self-training is a very basic type of semi-supervised learning
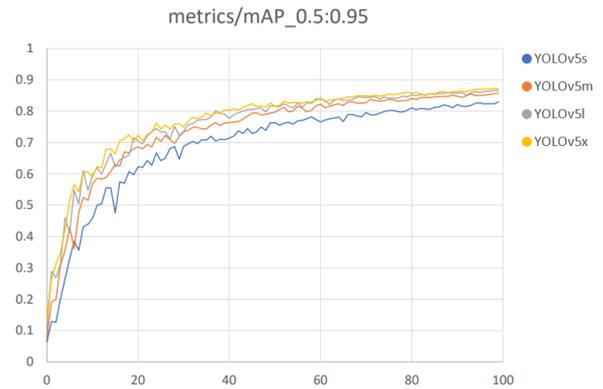
[27]. This method involves teaching a model on a tiny number of labeled data, then taking the data which the model is most confident about and creating a new model using both the original set and the one the model has labeled itself. This process is repeated iteratively until no further significant improvement is visible.

As a reference point for our experiments with active learning, we trained a baseline model using a conventional approach and employing the entire manually labeled dataset, without the assistance of active learning techniques. For this baseline model, we primarily utilized a confusion matrix to evaluate its performance in classifying various types of surgical instruments, focusing on precision and recall for each category. This detailed, category-specific analysis is crucial for understanding the model's strengths and weaknesses in classifying manually labeled images. The performance of this baseline model, including its confusion matrix, is illustrated in the figure 4.

In contrast, for other experiments involving active learning, we employed the mean Average Precision (mAP) metric. This choice was driven by mAP's capability to provide a broader overview of model performance across different classes. It is particularly effective for comparing the efficacy of various active learning strategies, offering a holistic view of model accuracy in a more dynamic learning context.

Four experiments were performed:

1. **High Confidence Self-Training Experiment:** In the first experiment, the self-training method was applied. Any detection with a confidence score higher than 0.95 was automatically assigned the predicted label and added to the dataset.

2. **Integrated Active Learning Experiment:** The second experiment, integrated the self-training method with the baseline active learning approach. Here, uncertainty sampling played a key role, with a specified confidence interval between 0.1 and 0.4. This specified threshold helped us to identify images that were problematic for our model, excluding any noise or arbitrary detections. Such images were then manually labeled. This strategy resulted in the additional annotation of 2110 images, of which 323 were manually anno-
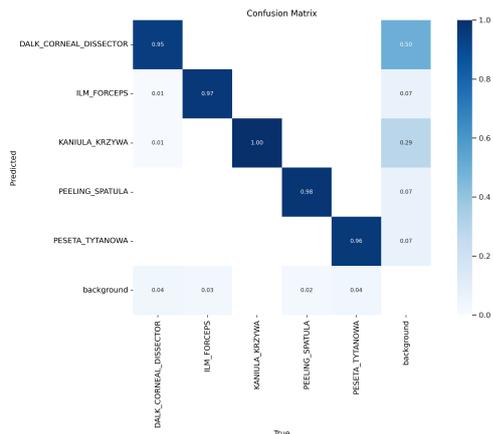
6

**Fig. 4.** Confusion Matrix of the Model Trained on Manually Labeled Images of Surgical Tools. This matrix depicts the model's ability to classify different types of surgical instruments, emphasizing its precision and recall for each category

tated and the remaining 1787 were annotated by automated processes.

3. **Subset Iterative Refinement Experiment:** The third experiment, integrated the self-training method, baseline active learning approach and batch split of images. Uncertainty sampling remained with confidence interval between 0.1 and 0.4. Additionally we implemented splitting the set of images we worked on to only 20% of the test-set and performed detection on it. The intention was to reduce the number of images labeled manually by doing iterations more often. If we had many similar images in the test set with which the model performs below expectations, it makes no sense to label each of these cases manually. Tagging just one of these images manually will improve the detection results for the remaining ones. During this experiment, we managed to additionally label 1486, of which 375 photos were manually labeled and 1111 were automatically labeled.

4. **Dynamic Confidence Threshold Experiment:**

   The fourth **experiment** is based on the second **experiment**, except that at each iteration, the threshold for which we labeled images automatically decreased by one percent. The idea behind this method is that, as the learning dataset grew larger and larger, the model no longer confused similar surgical instruments as often. It has learned to distinguish details better, meaning the model can be trusted more. So you can reduce its automatic approval threshold to allow more images to go through to the next iteration.

   During the experiment, we managed to additionally label 2510, of which 268 photos were manually labeled and 2242 automatically.

Table 3 presents the results of the experiments. In the table, the designation "(+696)" means that in each experiment, 20% of the set was manually labeled at the beginning, of which 10% were training images and 10% were verification images, accounting for 696 manually labeled images.

In summary, experiment number two achieved the best performance in the confusion matrix, closely followed by exper-

iment number four. The results achieved for the other two methods were slightly worse, with the third experiment taking a great many iterations.

As one can see, the fourth experiment succeeded in labeling the largest number of photos automatically, with a very small number of photos additionally marked manually. Experiment one and experiment four are the most interesting cases and are worth paying the most attention.

Although we did not give a single photo to be labeled by a human, the first experiment allowed additional labeling of 1,689 photos automatically, which is almost half of the available dataset in this case. This approach has its pros and cons. On one hand, it may overlook photos that could be highly relevant to the model since automatic tagging is limited to those with the fewest errors above the 95% confidence threshold. However, on the other hand, it enables us to efficiently label a vast number of images quickly. So does experiment number four, and it does it even faster. The fourth experiment also involved the automatic labeling of images above a certain threshold, with the difference being that with each successive iteration, we trusted the model more and more. In the beginning, we assumed that the confidence threshold for the automatic labeling of a photo was 95%, and with each iteration, it is dropped by one percent. This allows us to automatically label a much larger amount of data, but unfortunately, it comes with the consequences of less accurate labeling. The boundaries of the bounding rectangle in such cases happen to be less precise than in the case of experiment one. This method can be applied in cases where a model can afford to make more errors than in medical applications. For example, to detect cars in order to study traffic volume.

Experiment two investigated the impact of active learning, specifically additional labeling of a small number of poorly recognized images, on the subsequent automatic labeling of the remaining photos. By manually labeling 323 photos (9.3% of the total set), we enabled the automatic labeling of an additional 98 photos.

In experiment three, the same approach as in experiment two was taken, but with a modification. Each iteration was divided into 20 percent batches, aiming to accelerate the automatic labeling process by frequently using smaller portions of manually labeled photos. Unfortunately, after 19 iterations, this expected behavior was not observed, and further iterations became infeasible due to the prolonged learning time of a single model.

At the end of the experiment, the model took about 4-5 hours to learn one iteration, and there was a frequent timeout error in the Colab notebook. This made it very time-consuming to continue the experiment, and after early results, approach did not promise satisfactory results.

**Analysis of the results is presented in Table 3, which summarises the quantitative ratio of manually labelled to automatically labelled images in each experiment, and in the more detailed tables: Table 4 for Experiment 1, Table 5 for Experiment 2, Table 6 for Experiment 3, and Table 7 for Experiment 4, it can be concluded that active learning methods show a significant effectiveness in accelerating the**

7

Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska

model training process with a limited amount of data. In particular, Experiment 4, the results of which are detailed in Table 7 and illustrated in Figure 5, shows the highest auto-labelling efficiency. Similarly, Experiment 2, shown in Table 5, demonstrates how strategic manual labelling of selected images can help to improve the classification efficiency of the model, confirming the usefulness of active involvement in the algorithm training process.

8

**Table 3.** Number of detected images in each experiment

| Exp no. | Auto. | Hand labeled | Ratio of hand/auto lables | Hand labeled percent | Auto labeled percent |
|---------|-------|--------------|---------------------------|----------------------|----------------------|
| Exp. 1 | 1689 | 0 (+696) | 41,2% | 20,0% | 48,6% |
| Exp. 2 | 1787 | 323 (+696) | 57,0% | 29,3% | 51,4% |
| Exp. 3 | 1111 | 375 (+696) | 96,4% | 30,8% | 32,0% |
| Exp. 4 | 2242 | 268 (+696) | 43,0% | 27,7% | 64,5% |

**Table 4.** Comparison of classification error for experiment number one

| Surgical tool | 1-st | 2-nd | 3-rd | 4-th | 5-th | 6-th | 7-th | 8-th | 9-th |
|---------------|------|------|------|------|------|------|------|------|------|
| DALK_CORNEAL_D. | 0.85 | 0.89 | 0.93 | 0.95 | 0.93 | 0.96 | 0.95 | 0.96 | 0.95 |
| ILM_FORCEPS | 0.84 | 0.88 | 0.90 | 0.96 | 0.94 | 0.95 | 0.95 | 0.94 | 0.96 |
| CURVED_CANNULA | 0.96 | 0.96 | 0.95 | 0.96 | 0.96 | 0.96 | 0.97 | 0.97 | 0.97 |
| PEELING_SPATULA | 0.88 | 0.89 | 0.88 | 0.91 | 0.92 | 0.90 | 0.90 | 0.92 | 0.89 |
| TITANIUM_FORCEPS | 0.98 | 0.97 | 0.99 | 0.99 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |

**Table 5.** Comparison of classification error for experiment number two

| Surgical tool | 1-st | 2-nd | 3-rd | 4-th | 5-th | 6-th | 7-th | 8-th | 9-th | 10-th | 11-th |
|---------------|------|------|------|------|------|------|------|------|------|-------|-------|
| DALK_CORNEAL_D. | 0.85 | 0.90 | 0.95 | 0.94 | 0.96 | 0.97 | 0.97 | 0.99 | 0.98 | 0.96 | 0.95 |
| ILM_FORCEPS | 0.84 | 0.91 | 0.95 | 0.96 | 0.94 | 0.97 | 0.96 | 0.95 | 0.96 | 0.96 | 0.95 |
| CURVED_CANNULA | 0.96 | 0.96 | 0.97 | 0.98 | 0.97 | 0.99 | 0.99 | 0.98 | 0.99 | 0.98 | 0.98 |
| PEELING_SPATULA | 0.88 | 0.93 | 0.93 | 0.97 | 0.94 | 0.98 | 0.97 | 0.97 | 0.97 | 0.96 | 0.97 |
| TITANIUM_FORCEPS | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 0.99 |

**Table 6.** Comparison of classification error for experiment number three

| Surgical tool | 1-st | 2-nd | 3-rd | 4-th | 5-th | 6-th | 7-th | 8-th | 9-th | 10-th | 11-th | 12-th | .. | 18-th | 19-th |
|---------------|------|------|------|------|------|------|------|------|------|-------|-------|-------|-----|-------|-------|
| DALK_CORNEAL_D. | 0.85 | 0.79 | 0.85 | 0.86 | 0.83 | 0.89 | 0.84 | 0.93 | 0.88 | 0.87 | 0.90 | 0.92 | .. | 0.91 | 0.95 |
| ILM_FORCEPS | 0.84 | 0.87 | 0.87 | 0.92 | 0.90 | 0.91 | 0.92 | 0.91 | 0.93 | 0.97 | 0.93 | 0.95 | .. | 0.95 | 0.93 |
| CURVED_CANNULA | 0.96 | 0.95 | 0.96 | 0.97 | 0.96 | 0.98 | 0.98 | 0.97 | 0.97 | 0.96 | 0.97 | 0.98 | ... | 0.98 | 0.98 |
| PEELING_SPATULA | 0.88 | 0.90 | 0.89 | 0.90 | 0.94 | 0.88 | 0.92 | 0.95 | 0.93 | 0.93 | 0.92 | 0.94 | ... | 0.93 | 0.93 |
| TITANIUM_FORCEPS | 0.98 | 0.95 | 0.97 | 0.96 | 0.98 | 0.99 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | ... | 0.99 | 1.00 |

**Table 7.** Comparison of classification error for experiment number four

| Surgical tool | 1-st | 2-nd | 3-rd | 4-th | 5-th | 6-th | 7-th |
|---------------|------|------|------|------|------|------|------|
| DALK_CORNEAL_D. | 0.85 | 0.90 | 0.96 | 0.93 | 0.96 | 0.95 | 0.96 |
| ILM_FORCEPS | 0.84 | 0.91 | 0.96 | 0.93 | 0.95 | 0.95 | 0.95 |
| CURVED_CANNULA | 0.96 | 0.96 | 0.97 | 0.98 | 0.97 | 0.99 | 0.98 |
| PEELING_SPATULA | 0.88 | 0.93 | 0.96 | 0.96 | 0.95 | 0.95 | 0.95 |
| TITANIUM_FORCEPS | 0.98 | 0.99 | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 |

Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska

**Table 8.** Summary of Experiments up to the 7th Iteration of Active Learning Training

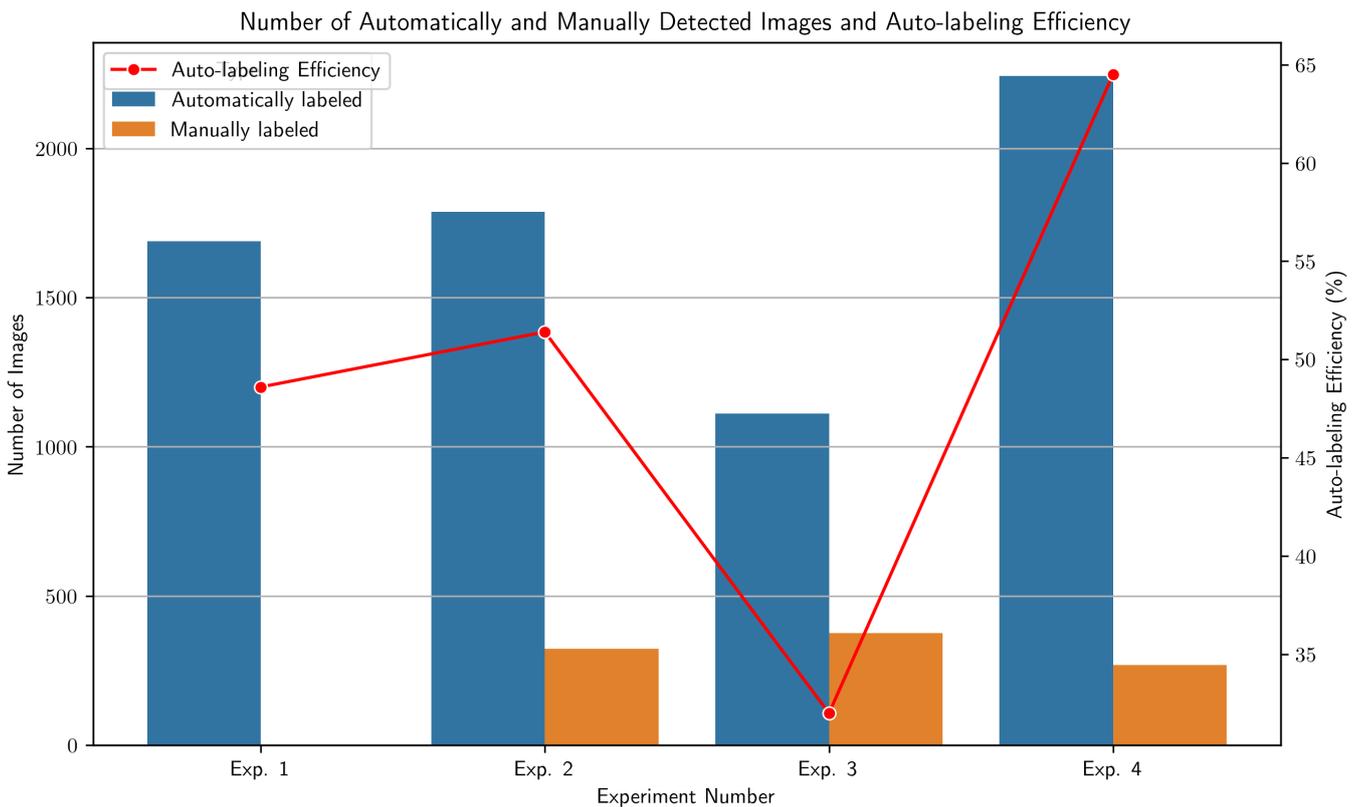| Tool | | | | | | Average |
|---|---|---|---|---|---|---|
| **Tool Tip** | | | | | | |
| **Experiment** | DALK_CORNEAL.D | ILM_FORCEPS | CURVED_CANNULA | PEELING_SPATULA | TITANIUM_FORCEPS | |
| 1 | 0.96 | 0.95 | 0.97 | 0.92 | 0.99 | 0.958 |
| 2 | 0.97 | 0.95 | 0.99 | 0.97 | 1.00 | 0.976 |
| 3 | 0.91 | 0.93 | 0.98 | 0.94 | 1.00 | 0.952 |
| 4 | 0.96 | 0.95 | 0.98 | 0.95 | 1.00 | 0.968 |



**Fig. 5.** Illustrates the ratio of manually to automatically labeled images across the experiments, with the red line denoting the trajectory of auto-labeling efficiency, peaking significantly in Experiment 4.

## 7. CONCLUSIONS AND FUTURE RESEARCH

In the rapidly advancing technology era, machine learning has become an indispensable tool in various domains, including medical imaging. Among the different machine learning approaches, active learning has garnered significant attention due to its potential to optimize the learning process while minimizing data acquisition costs. This is particularly useful when dealing with a limited dataset, such as medical images.

However, it is essential to thoroughly investigate the trade-offs associated with active learning compared to classic supervised learning methods, particularly concerning performance effects. Our research focuses on exploring the usefulness of active learning in the context of surgical image detection. While active learning offers the advantage of utilizing a smaller training set, there may be compromises in achieving optimal results compared to traditional supervised learning. We have undertaken a novel approach to address this challenge by integrating semi-supervised learning techniques to improve the model. Furthermore, we recognize the significance of the initial selection of the training data in the model's performance. The set is currently chosen randomly, but we believe that incorporating clustering methods can lead to more diverse and informative training data and improve the model's generalization ability. In the subsequent stages of our research, we delve into the specific task of surgical tool localization within medical images. The accurate detection and precise positioning of surgical instruments are pivotal in ensuring successful surgical outcomes. To achieve this, we plan to apply classical image processing methods after detecting the relevant region in the image containing the surgical tool. This two-step approach enables us to achieve both efficiency and accuracy in the tool localization process.

## ACKNOWLEDGMENTS

## REFERENCES

[1] D. Zhou, S. Kimura, H. Takeyama, D. Haraguchi, Y. Kaizu, S. Nakao, K. Sonoda, and K. Tadano, "Eye explorer: A robotic endoscope holder for eye surgery," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 17, 10 2020.

[2] B. Becker and C. Riviere, "Real-time retinal vessel mapping and localization for intraocular surgery," *IEEE International Conference on Robotics and Automation : ICRA : [proceedings] IEEE International Conference on Robotics and Automation*, pp. 5360–5365, 05 2013.

[3] M. Rosenfield and N. Logan, *Optometry: Science, Techniques and Clinical Management*. Elsevier Ltd, 2009.

[4] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[5] M. Alsheakhali, M. Yigitsoy, A. Eslami, and N. Navab, "Surgical tool detection and tracking in retinal microsurgery," in *Proceedings of SPIE Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2015, pp. 1–4, event: SPIE Medical Imaging, 2015. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie

[6] K. Gromada, B. Piotrowski, A. Ciąćka, P.and Kurek, and A. Curatolo, "Improved tool tracking algorithm for eye surgery based on combined color space masks," in *Proceedings of SPIE Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling*, 2023, p. 124660G, event: SPIE Medical Imaging, 2023, San Diego, California, United States. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie

[7] C. Lin, Y. Zheng, C. Guang, K. Ma, and Y. Yang, "Precision forceps tracking and localisation using a kalman filter for continuous curvilinear capsulorhexis," *Robotics Computer Surgery*, vol. 18, p. e2432, 2022.

[8] G. Luijten, C. Gsaxner, J. Li, A. Pepe, N. Ambigapathy, M. Kim, X. Chen, J. Kleesiek, F. Hölzle, B. Puladi, and J. Egger, "3d surgical instrument collection for computer vision and extended reality," *Sci Data*, vol. 10, p. 796, 2023.

[9] M. Allan, S. Ourselin, S. Thompson, D. Hawkes, J. Kelly, and D. Stoyanov, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering*, vol. 60, pp. 1050–1058, 2013.

[10] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin, "Detecting surgical tools by modelling local appearance and global shape," *Trans. on Medical Imaging*, vol. 34, no. 12, 2015.

[11] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical Image Analysis*, vol. 35, 2017.

[12] J. Zhou and S. Payandeh, "Visual tracking of laparoscopic instruments," *J. Autom. Cont. Eng.*, vol. 2, no. 3, pp. 234–241, 2014.

[13] R. Sznitman, R. Richa, R. Taylor, B. Jedynak, and G. Hager, "Unified detection and tracking of instruments during retinal microsurgery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, 2013.

[14] N. Rieke, D. Tan, C. Amat di San Filippo, F. Tombari, M. Alsheakhali, V. Belagiannis, A. Eslami, and N. Navab, "Real-time localization of articulated surgical instruments in retinal microsurgery," *Medical Image Analysis*, vol. 34, 2016.

[15] N. Rieke, D. Tan, F. Tombari, J. Vizcaíno, C. di San Filippo, A. Eslami, and N. Navab, "Real-time online adaption for robust instrument tracking and pose estimation," in *MICCAI*. Springer, 2016, pp. 422–430.

[16] X. Yang, Y. Zhang, and D. Zhou, "Deep networks for image super-resolution using hierarchical features,"

Bartłomiej Piotrowski, Jakub Oszczak, Krzysztof Sawicki, Barbara Siemiątkowska

*Bulletin of the Polish Academy of Sciences Technical Sciences*, vol. 70, no. 1, p. e139616, 2022. [Online]. Available: http://journals.pan.pl/Content/121550/PDF-MASTER/1949_corr.pdf

[17] B. Y. Suprapto, Kelvin, M. A. Kurniawan, M. K. Ardela, H. Hikmarika, Z. Husin, and S. Dwijayanti, "Identification of garbage in the river based on the yolo algorithm," *International Journal of Electronics and Telecommunications*, vol. vol. 67, no. No 4, pp. 727–733, 2021. [Online]. Available: http://journals.pan.pl/Content/121912/PDF-MASTER/100_3225_Suprapto_skl1.pdf

[18] T. e. a. Mahendrakar, "Performance study of yolov5 and faster r-cnn for autonomous navigation around non-cooperative targets," *2022 IEEE Aerospace Conference (AERO)*, 2022, https://arxiv.org/ftp/arxiv/papers/2301/2301.09056.pdf, access 10.03.2023.

[19] B. Settles, *Active Learning*. Morgan Claypool Publishers, 2012.

[20] F. Marquardt, "Lecture 26: Active learning for network training: Uncertainty sampling and other approaches." https://www.youtube.com/watch?v=fwHZtqr-uBY, access 01.03.2023.

[21] B. Zhang, L. Li, S. Yang, S. Wang, Z.-J. Zha, and Q. Huang, "State-relabeling adversarial active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8756–8765.

[22] O. M. Cliff, M. Prokopenko, and R. Fitch, "Minimising the kullback–leibler divergence for model selection in distributed nonlinear systems," *Entropy*, vol. 20, no. 2, 2018. [Online]. Available: https://www.mdpi.com/1099-4300/20/2/51

[23] L. e. a. Wang, "Active learning via query synthesis and nearest neighbour search," *Neurocomputing*, no. 147, pp. 426–434, 5 2015, https://www.sciencedirect.com/science/article/abs/pii/S0925231214008145, access 01.03.2023.

[24] K. Lang and E. Baum, "Query learning can work poorly when a human oracle is used." *IEEE Press*, pp. 335–340, 1992, https://www.academia.edu/6168656/Query_learning_can_work_poorly_when_a_human_oracle_is_used, access 01.03.2023.

[25] K. e. a. He, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Lecture Notes in Computer Science*, p. 346–361, 2014, https://arxiv.org/pdf/1406.4729.pdf, access 10.03.2023.

[26] Z. e. a. Zheng, "Distance-iou loss: Faster and better learning for bounding box regression," *AAAI*, 2020, https://arxiv.org/pdf/1911.08287.pdf, access 10.03.2023.

[27] O. Chapelle, B. Scholkopf, and A. Zien, *Semi-Supervised Learning*. The MIT Press, 09 2006. [Online]. Available: https://doi.org/10.7551/mitpress/9780262033589.001.0001

12