Robust Semi-supervised Detection of Hands in Diverse Open Surgery Environments

Pranav Vaid PVAID@STANFORD.EDU
Department of Computer Science
Stanford University
Stanford, California, USA

Serena Yeung SYYEUNG@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, California, USA

Anita Rau ARAU@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, California, USA

Abstract

Artificial intelligence has impacted many aspects of modern medical care but depends critically on data. Videos of medical procedures are a valuable resource for computer vision algorithms but labeling them can be costly and requires expert knowledge. This paper explores how to leverage low-quality, unlabeled videos scraped from the internet in addition to a limited amount of labeled images to improve object detection during surgical procedures. We establish the first benchmark for semi-supervised hand detection during open surgery and show that existing benchmarks in non-medical contexts are not indicative of performance differences on real-world medical applications, where data is noisy and poorly labeled. We propose a end-to-end trainable two-stage object detector that employs consistency loss to learn from unlabeled images. The model is robust to missing labels, variance in hand morphology, and extreme domain shifts such as those encountered in open-source videos of surgeries scraped from YouTube. Our method can predict surgeons' hands in surgical videos even when only a fraction of hands are labeled in each frame of the labeled set. Adding unlabeled data, we can detect hands more accurately than existing end-to-end semi-supervised object detection algorithms.

1. Introduction

Deep learning has made impressive strides in recent years, achieving unparalleled accuracy and reliability across many tasks. These algorithms owe their success partially to the availability of large labeled datasets that enable networks to generalize to previously unseen examples. However, the medical field presents unique challenges to collecting large labeled datasets, including privacy concerns, the need for expert labelers, and the lack of data collection tools. Learning algorithms developed in general scenes are thus not always directly applicable to the medical domain, especially for complex computer vision tasks. Fortunately, while labeled medical datasets are difficult to obtain, there is a wealth of unlabeled medical

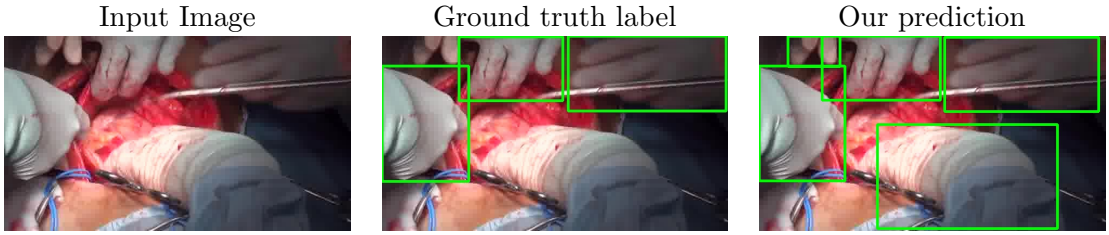

Figure 1: Even the small subset of labeled images in AVOS is incomplete. The ground truth labels miss obvious hands that our model detects.

data that is easily accessible. For example, surgery videos are publicly available on YouTube, providing a valuable resource for medical research.

Surgeries are critical procedures for treating various medical conditions, but they also carry significant risks of complications. Improving the safety and efficiency of surgical procedures is therefore essential. One promising approach is automated surgical understanding and analysis, which can provide valuable feedback to surgeons during operations. But research efforts have mainly focused on laparoscopic surgeries because they provide video data by default, making minimally invasive procedures more amenable to deep learning approaches. However, many surgical procedures cannot be completed using minimally invasive techniques. These procedures include organ transplantations, open heart surgery, mastectomies, hip and knee replacements, and plastic surgeries. Despite the importance and prevalence of open and surface surgery, deep learning methods for surgical understanding in these settings are significantly less developed than in laparoscopy.

To bridge the research gap between laparoscopy and open surgery in computer vision, we explore the detection of hands in open surgery videos. In these procedures, surgeons directly manipulate tissues and instruments with their hands. Their detection can facilitate downstream tasks such as action recognition, surgical skill prediction, or precise hand pose estimation. We propose to mitigate the lack of standardized, labeled data for open surgeries with semi-supervised learning techniques that leverage small numbers of incomplete labels and fully unlabeled publicly available data. Hands are especially difficult to detect in a surgical context since they morph appearance often, and vary significantly in their color and texture due to the use of gloves or other protective equipment. We use the AVOS dataset (Goodman et al. (2021)), which contains 1,997 diverse open surgery videos scraped from YouTube and associated hand bounding boxes in selected frames from these videos. This dataset is one of the only available containing bounding boxes for surgical hands. To the best of our knowledge the only other similar dataset is Louis et al. (2023) which contains only 28 surgical videos. Although AVOS provides substantially more videos, the dataset presents its own set of challenges. As the videos are scraped from YouTube, the filming is not standardized and videos vary significantly in their quality, lighting, camera pose, and occlusion. The surgeries captured in the dataset are also diverse, representing 23 different procedures from 50 countries. Most of the dataset remains unlabeled: out of the entire dataset, only 334 videos contain labels with hands. Roughly ten frames are labeled per video, but even in labeled frames many hands were missed by the labelers (see Figure 1).

This paper explores how this easily accessible, but unlabeled and heterogeneous data can be leveraged to enhance the performance of deep learning models on real-world medical procedures. Our contributions are threefold:

- To the best of our knowledge, we establish the first benchmark for hand detection during surgery. We perform a comprehensive study of performance of current state of semi-supervised learning approaches on the hand detection task in open surgery.
- We explore two major approaches, STAC and Soft Teacher, and explore why these state-of-the-art approaches can be brittle to the properties of surgical image data.
- Based off these findings, we introduce a novel hand detection model that outperforms existing methods while being end-to-end trainable.

To facilitate future research, we make our model and trained weights publicly available[1].

Generalizable Insights about Machine Learning in the Context of Healthcare

Our work provides several generalizable insights: Firstly, we show that unlabeled low-quality data scraped from the internet can help improve the performance of object detection networks in the medical domain. Rather than relying on expensive expert labelling, exploring more unconventional data sources can be useful. Secondly, we establish the first surgery hand detection benchmark for semi-supervised methods and show that comparisons of models on existing benchmarks are not always transferable to real-world medical applications. In some cases, best performing methods on urban, outdoor, and indoor environments fail at outperforming basic baseline models on surgical scenes, which can be extremely heterogeneous and noisy. Using alleged *best* methods blindly can therefore lead to subpar results. Thirdly, we show that some properties of surgical scenes can be used to our advantage. While researchers usually focus on *mitigating* effects of surgical video, we leverage the fact that surgeries cannot be performed by arbitrary numbers of surgeons, allowing us to apply methods which would otherwise not be possible. As data in healthcare often has unique properties, it is important to understand and leverage those in an advantageous manner for medical-specific machine learning approaches. Lastly, our new method is applicable to the detection of other objects in surgical video, like certain anatomies, tools, or other items used during surgery. Our model is robust towards poor data quality and incomplete labels, making it applicable to many different data availability scenarios.

2. Related Work

Several well established works study the use of semi-supervised learning to leverage unlabeled data for object detection. We review the current state of semi-supervised learning techniques for object detection and their application to medical imaging.

Consistency based. Semi-supervised methods incorporate losses derived from both labeled and unlabeled data. While supervision from labeled images is straight-forward, there are different approaches to learning from unlabeled images. One common approach are consistency-based approaches, which encourage models to make consistent predictions on

1. https://github.com/pranavvaid/robust-csd

augmented or noisy versions of the same unlabeled image. Tang et al. (2021) use a consistency loss to teach their model to learn noise-robust proposal features from labeled and unlabeled data by adding noise to the convolutional feature maps. Jeong et al. (2019) augment images directly with flip augmentations and enforce consistency between proposals in the unaugmented and augmented versions of the same image. However, one drawback of their approach is that matching the high quantity of proposals is intractable for two stage detectors, forcing the authors to use a consistency localization loss with only one stage detectors, which generally perform worse. We use surgical domain knowledge to address this issue, allowing a localization loss to be utilized with two stage detectors in our model.

Pseudo labels. Another well established approach for semi-supervised object detection is the use of pseudo labels, in which unlabeled images are annotated with predictions from an initially trained teacher model, and then used to train a final student model. Works such as Radosavovic et al. (2018) build on the traditional pseudo labeling scheme by ensembling predictions of different data augmentations to generate pseudo labels. More recent works such as STAC (Sohn et al. (2020)), achieve significantly better results on benchmark datasets like MS-COCO (Lin et al. (2015)) and PASCAL VOC (Everingham et al. (2010)) by utilizing consistency regularization on pseudo labels of strongly augmented unlabeled train images. STAC, like many other pseudo labeling schemes, requires a multi-stage training scheme. Recent works such as Xu et al. (2021) propose Soft Teacher, an end-to-end pseudo labeling framework that updates the teacher model throughout the training process alongside the student model with an exponential moving average strategy. Soft Teacher maintains a simpler training process than STAC, while also outperforming it on the COCO dataset.

Application to health care. Recently, a growing body of work surrounds the use of semi-supervised learning in the medical imaging domain. Wei et al. (2022) and Wu et al. (2022) propose teacher-student mutual learning frameworks for object detection of femur fractures and retinal lesions, respectively, on images of medical scans such as X-Rays and OCT B scans. Zhou et al. (2021) add an adapted consistency loss to the one-stage detector RetinaNet (Lin et al. (2018)) for nuclei detection of cells. Semi-supervised methods have been shown to help performance of detectors in these domains, although the challenges associated with object detection in medical scans differ sharply from the challenges associated with the surgical image domain, which include highly variable recording equipment, shifting point of view, and heterogeneous objects.

Research on semi-supervised techniques for surgery has mostly focused on minimally invasive surgeries. Jiang et al. (2021), Yoon et al. (2020), and Teevno et al. (2022) utilize supervised frameworks to improve detection of surgical tools in minimally invasive surgeries. Our task of hand detection in open surgery images differs from these works in two main ways. Firstly, tools in minimally invasive surgeries have a more morphologically stable appearance, and thus do not present the same challenges as detection of deformable objects like hands in medical images. Secondly, tool detection is a multiclass problem and semi-supervised frameworks used by these works are optimized to increase classification accuracy in spite of issues like class imbalance. We are only concerned with the single class problem of hand detection, and need semi-supervised frameworks optimized for improving localization accuracy.

Hand detection in an open surgery context remains largely unexplored. Zhang et al. (2020) explore detection of hands from online YouTube videos using supervised models with

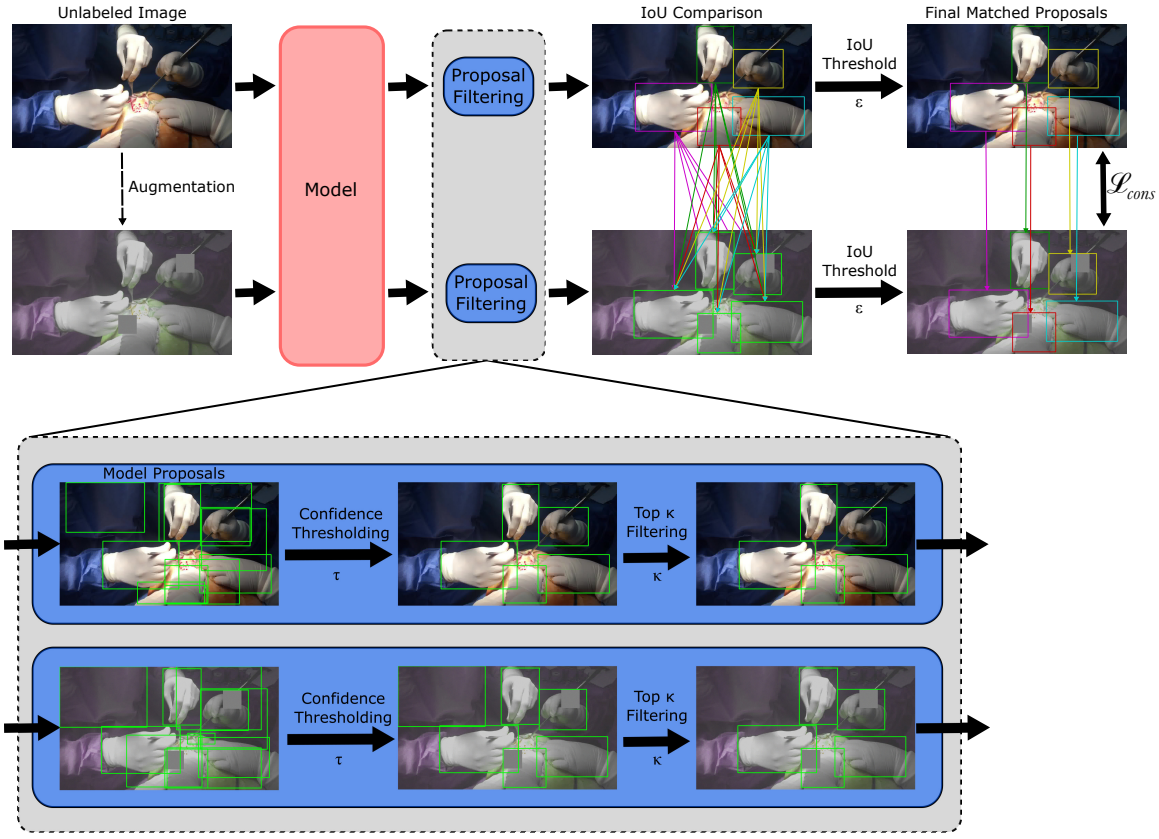

Figure 2: Overview of our method leveraging unlabeled images. Unlabeled images are passed through the object detection model. Proposals are then filtered in a multi-step approach, and lastly matched.

pretraining. Goodman et al. (2021) utilize a hand detection model in open surgery videos to track hand movements, detect keypoints, and predict surgeon skill. None of these works have explored the application of semi-supervised frameworks to improve performance on this task. Furthermore, while approaches such as STAC and Soft Teacher have been shown to increase performance on benchmark datasets such as MS-COCO (Lin et al. (2015)), real world medical datasets such as the AVOS surgery dataset presents challenges not present in standardized benchmark datasets such as varied lighting and quality, highly deformable objects, and inconsistent labeling. The performance of state-of-the-art semi-supervised algorithms in this setting, with these unique properties, is yet to be explored.

3. Methods

We propose an end-to-end trainable object detection framework based on consistency loss. Leveraging surgery-specific domain knowledge, we integrate our localization consistency loss into a two-stage detector, which was not previously possible.

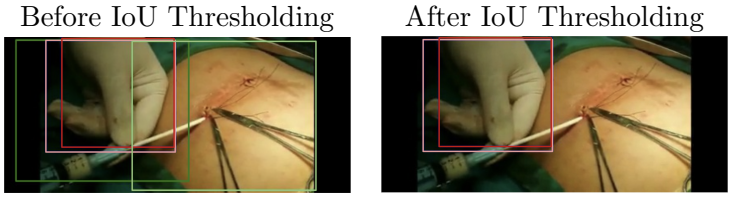

Figure 3: Requiring a minimum IoU threshold filters out incorrectly matched proposals. The lighter colored boxes represent predictions made on the augmented version of the image, the darker colored boxes represent the predictions made on the unaugmented version. The two pink boxes and two green boxes have been matched because they have the highest corresponding IoU, but after applying IoU thresholding the green pairing of proposals are discarded.

3.1. Consistency Loss for Object Detection

Two-stage detectors are generally more accurate than single stage object detectors (Sultana et al. (2020)). In the first stage, a region proposal network (RPN) finds regions of the image likely to contain objects. In the second stage, a regional convolutional neural net (R-CNN) predicts the classification scores and regional offset of the proposals from the RPN.

Consistency losses for object detection require matching proposals between unaugmented and augmented versions of images. This task is trivial for single stage detectors, as they are comprised of a single network whose output feature map correlates directly to the image input: consequently each pixel of the feature map should output predictions over the same objects, and it is trivial to match proposals. However, two stage detectors utilize a region proposal network that can output differing box proposals in the augmented and unaugmented image, such that all proposals in both images would have to be matched. Solving the correspondence matching problem is challenging and computationally expensive. Past works exploring consistency loss, such as Jeong et al. (2019), have therefore been unable to utilize a consistency localization loss with two stage detectors.

We propose a simple framework for consistency loss based on domain knowledge of the surgical domain that allows for us to avoid this issue, and utilize a consistency localization loss with two stage detectors. Although in traditional object detection tasks, there may be no upper limit on how many objects we may expect to detect in an image, we know that in an open surgery there will rarely be more than a couple operating surgeons, and thus we know that there will always be a low prevalence of hands or other surgical objects in any given frame of a surgical video. By incorporating this domain knowledge, we are able to develop a procedure to limit the amount of proposals that allows the correspondence matching problem to become computationally inexpensive. Bypassing the correspondence matching problem also allows us to utilize surgery specific augmentations such as color and quality based augmentations, rather than only simple flipping augmentations, which better reflect the variance in video quality and appearance of surgical videos.

3.2. Adapted Consistency Loss for Surgical Domain

The steps to our procedure are summarized as follows, and depicted in Fig. 2:

1. **Filter proposals** by retaining the most relevant predictions.
2. **Match proposals** by maximizing the Intersection over Union (IoU) of the proposal pairings in the original and augmented image.
3. **Compute localization loss** using the matched proposals.

Filter proposals. In order to make the correspondence matching problem tractable, it is necessary to reduce the amount of proposals in the original and augmented image that must be matched. An overview of our proposal filtering strategy is shown in Figure 2. We first apply a confidence threshold τ to the proposals on the regular and augmented images, in order to only retain predictions the model is confident are hands. However, with the use of a low confidence threshold, there will likely still be many proposals remaining after this step. Open surgeries never involve more than a few surgeons actively operating on a patient at any given time, which places an upper limit on the amount of hands we're likely to find in any given frame of a video. Thus, we also then apply a top κ filter, where we only retain the κ most confident proposals, in order to place a hard limit on the number of proposals we will consider as relevant. κ represents the upper limit of the objects of relevance we expect to see in any image within our domain. Thus, at maximum we must calculate κ^2 potential matches of proposals per image.

Match proposals. To match proposals between the regular and augmented image, we calculate the IoU between each potential pair of proposals in the image pair. We assign each proposal in the original image to the proposal in the augmented image with the highest corresponding IoU. Formally, this pairing is defined by the following, where p_i represents the i-th proposal in the regular image and $p_j^{'}$ represents the j-th proposal in the augmented image:

$$\max \texttt{IoU}(p_i, p_j^{'}); 0 \leq i, j < p$$

It is possible that after this step, proposal pairings are generated on entirely different objects due to a misprediction in either the regular or augmented image. These pairings would incorrectly produce an extremely high displacement, resulting in a high localization loss, which may affect model performance negatively. To counteract this, we also apply a minimum IoU threshold ϵ, where any matched pairing with an IoU below ϵ is ignored. An example of this is illustrated in Figure 3.

Compute localization loss. At this stage, each bounding box pair across the unaugmented and augmented image should represent the same object, and thus should have equivalent bounding boxes. To encourage the model to produce equivalent predictions, we compute a consistency localization loss over the two boxes. We can represent the bounding box of a prediction for any image as $[c_x, c_y, c_w, c_h]$, where c_x and c_y represent the center coefficients of the candidate box, and c_w and c_h represent the scale coefficients. With this, we can represent the predictions of a bounding box pairing (b, b') where b represents the bounding box of a prediction on the unaugmented image and b' represents the bounding box of the corresponding prediction on the augmented image, as $b = [c_x, c_y, c_w, c_h]$ and $b' = [c'_x, c'_y, c'_w, c'_h]$.

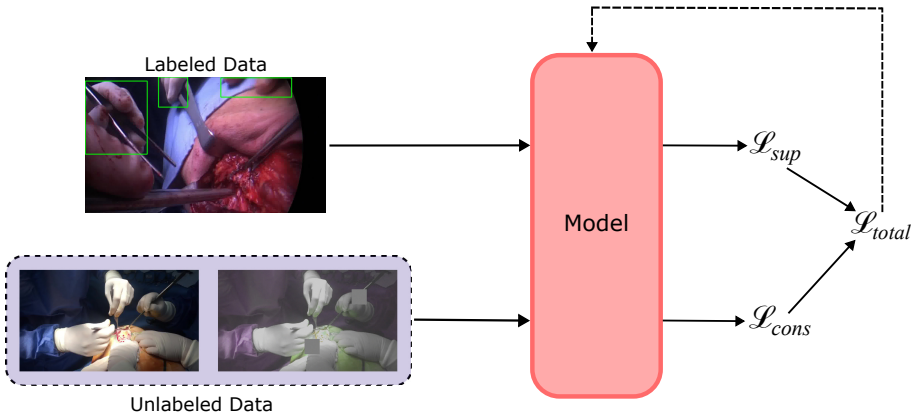

Figure 4: Our training setup. At train time, both labeled and unlabeled data are passed through the object detection model. We compute the supervised loss for the labeled data and consistency loss for the unlabeled image pairs.

For each matched bounding box pair, we then compute a consistency localization loss from the following displacement equation:

$$l_{cons} = \frac{1}{4}(\|c_x - c'_x\|^2 + \|c_y - c'_y\|^2 + \|c_w - c'_w\|^2 + \|c_h - c'_h\|^2)$$

The total consistency localization loss for an image is the average of the individual loss of each bounding box pair (b_p, b'_p), for a total localization loss of:

$$\mathcal{L}_{cons} = \mathbb{E}_p[l_{cons}(b_p, b'_p)]$$

Given that often times in the medical domain, the amount of labeled data is extremely low but the amount of unlabeled data can be high, it may be desirable to sample multiple unlabeled images when computing the consistency loss. Given the n sampled unlabeled images per labeled image, this is simply done by taking the summation of the consistency loss for each of the sampled images i as follows:

$$\mathcal{L}_{cons\_total} = \sum_{i=1}^{n} \mathcal{L}_{cons}^{(i)}$$

The final loss for the model is calculated as the supervised loss, summed with the weighted consistency, where λ_u represents the weight of the consistency loss.

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda_u \cdot \mathcal{L}_{cons\_total}$$

The full training pipeline is shown in Figure 4.

4. Experiments

In this section, we first establish a benchmark on hand detection in open surgery video. We compare two semi-supervised state-of-the-art methods, Soft Teacher and STAC, and evaluate them against their supervised baseline. We also train and evaluate a fully supervised

	labeled	unlabeled	**AP@50 ↑**	**AP@75 ↑**
Supervised baseline	✓		86.6	59.7
Supervised baseline w/ augment.	✓		88.3	64.0
Soft Teacher	✓	✓	86.6	51.3
STAC	✓	✓	87.9	**65.7**
Ours	✓	✓	**88.8**	62.6

Table 1: Comparison of different semi-supervised and fully supervised object detectors on the AVOS hand dataset. Bold numbers indicate best performance, while red methods highlight when model do not outperform the supervised baseline.

	True Positives @50 ↑	**False Negatives @50 ↓**
Supervised baseline	**920 (89.5%)**	**108 (10.5%)**
Supervised baseline w/ augment.	887 (86.3%)	141 (13.7%)
Soft Teacher	878 (85.4%)	150 (14.6%)
STAC	914 (88.9%)	114 (11.1%)
Ours	<u>917</u> (<u>89.2%</u>)	<u>111</u> (<u>10.8%</u>)

Table 2: True positives and false negatives for images with an IoU of at least 50%. True positive and false negative rates are shown in parentheses. We do not report false positives, as the dataset misses many ground truth hands, making false positives an unreliable measure. Bold numbers indicate the best performing model, while underlined numbers highlight the runner-up.

method, that is carefully augmented during training. All four models are compared to our proposed method in several data settings.

First, we use all available labeled and unlabeled images and show that our method can leverage unlabeled data and detect hands more accurately than the semi-supervised baselines. Second, we show that semi-supervised methods are especially useful in low-data regimes, while augmentation alone can be on-par when enough training data is available. Thirdly, we demonstrate that our method can handle incomplete labels more robustly than other semi-supervised methods.

4.1. Implementation details, data, and evaluation metrics

We use the Faster R-CNN framework (Ren et al. (2016)) implemented in Detectron2 Wu et al. (2019) as backbone for our model. STAC and Soft Teacher are both built on the R-CNN framework as well, allowing comparability between the methods, and making the vanilla Faster R-CNN a natural choice for our supervised baseline. To improve model ability to detect hands, all methods include a ResNet-50 backbone with weights pretrained on COCO (Lin et al. (2015)) and the EgoHands dataset (Bambach et al. (2015)), which contains labeled data of hands from first-person interaction between two people.

We use the default parameters and augmentations of Soft Teacher (confidence threshold of 0.9) and STAC ($\tau = 0.9$ and $\lambda_u = 2$) to train the baselines. For our model, we find that an output confidence threshold of $\tau = 0.9$, a top κ filter of $\kappa = 5$ proposals, and an IoU threshold of 0.25 provide us with the best performance on this task. Our model utilizes less extreme augmentations than Soft Teacher and STAC, as extremely augmented images can prevent or severely affect model predictions, preventing the localization loss from being useful. We apply all of the following augmentations to each unlabeled image in a random order: Additive Gaussian Noise (with mean of 0, scale of [0, 12.75]), Hue/Saturation modification (with value [-50, 50]), Linear Contrast (with alpha [0.5, 2]), and Cutouts (with 1-2 iterations at size 0.15). For our experiments, we set a sample ratio of $n = 8$ and an unsupervised consistency loss weight of $\lambda_u = 10$. All models use a resize operation on the labeled images during training, but for our best performing model we do not apply a resize operation to the unlabeled data. Our implementations of STAC and our model are built on the Detectron2 training framework. AVOS contains many spatial and temporal labels for the surgical videos, but for our investigation we only utilize the labeled videos and annotations containing hands. Our train set contains 240 labeled videos, validation set contains 40 videos, and test set 54 videos. This results in 1,941 labeled frames in the train set (with 4,473 annotations of hands), 325 labeled frames in the validation set (with 722 annotations of hands), and 457 labeled frames in the test set (with 1,028 annotations of hands). We uniformly sample 10 frames from 1,634 unlabeled videos, generating 16,340 unlabeled frames that we use as our unlabeled training dataset.

Our main metric is the average precision (AP), which represents the area under the precision-recall curve at a particular IoU threshold. In our experiments we use the thresholds 50 and 75. When models predict several hand bounding boxes within each other, we only assign one hand as true positive match, namely the one with the highest IoU.

Our secondary evaluation metric are the numbers of true positive and false negative hands. Similarly to the IoU, when several hands are predicted within one another, we assign ground truths to the prediction with the highest IoU only. Unassigned predictions are considered to be false positives, and ground truths with no corresponding predictions above the IoU threshold are considered false negatives.

4.2. Semi-supervised hand detection

First, we evaluate how well the different models leverage unlabeled data. We train each semi-supervised model according to their training protocols on the entire labeled and unlabeled datasets. The baseline is trained on the labeled data only. We also augment the baseline with our augmentations to separate the effect of the augmentations form the use of unsupervised data. Table 1 shows an overview of the performance of each model. Surprisingly, Soft Teacher fails to outperform the unaugmented baseline on both metrics, indicating that adding unlabeled data according to the method's protocol hurts performance rather than improving it. Our method, on the other hand, improves the baseline by 2.2 percentage points on the AP @ 50, and 2.9 points on the AP @ 75. Our method outperforms STAC at AP @ 50 by 0.9 points, but underperforms STAC on the AP @ 75 by 3.1 points. At the high level, this suggests our method underperforms STAC at precisely localizing hands, but is better at detecting hands given looser localization constraints. For the downstream

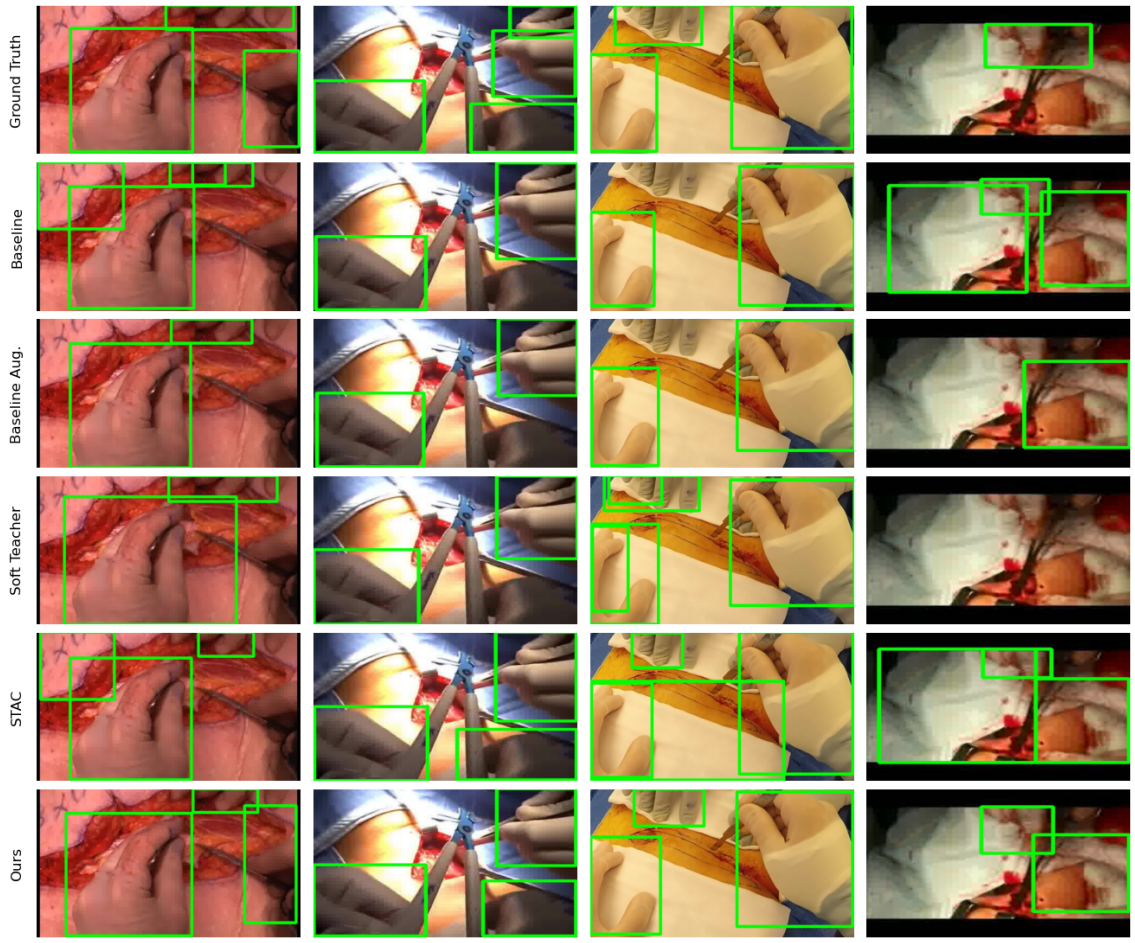

Figure 5: Comparison of model predictions with minimum confidence score of 0.5. While all methods predict obvious hands successfully, not all methods detect hard-to-spot ones. Note that the last column misses a hand in the ground truth label.

tasks most relevant to this problem, such as action recognition and skill prediction, it might be more useful to be able to better detect as many of the present hands as possible with a reasonable location rather than a few of the hands with a very precise location. STAC outperforms the baseline, however it does not outperform the augmented baseline on the AP @ 50. Just augmenting the labeled images is therefore a valid alternative to state-of-the-art semi-supervised methods. We posit that the heterogeneous open surgery dataset does not allow Soft Teacher and STAC to leverage the unlabeled data, as the generated pseudo labels are too inaccurate to be useful.

To understand the predictions better, we divide them into true positives and false negatives. As the dataset is poorly labeled, we do not show false positives, as these examples could indeed be true positives with wrong ground truth. Table 2 compares true positives and false negatives across all methods. Most notably, the supervised baseline predicts the

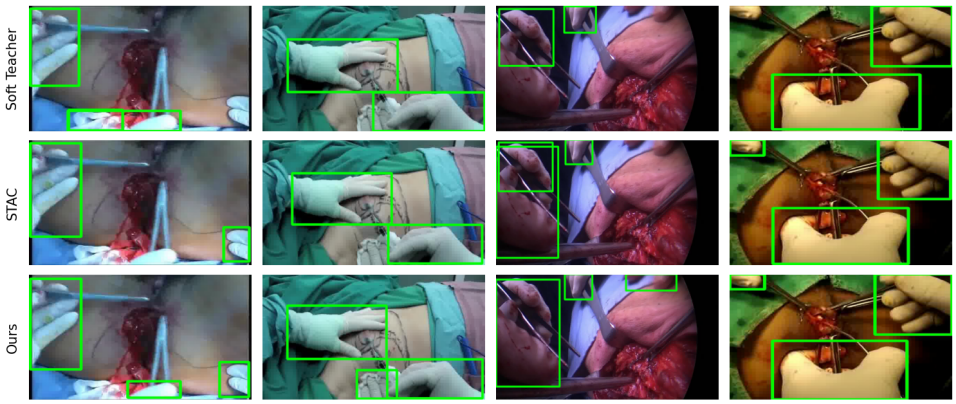

Figure 6: Our method detects partials hands more often than the baseline models.

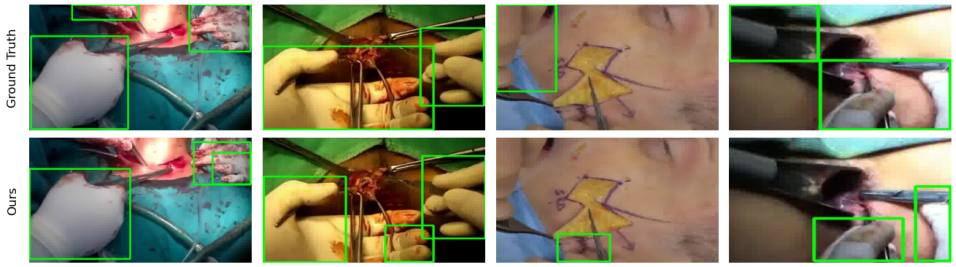

Figure 7: Failure cases of our method.

least false negatives, closely followed by our method. That is, the supervised baseline misses least hands of those that are labeled. But investigating the qualitative examples in Figure 5, we can observe that the supervised baseline over-predicts hands. It predicts several hands where only one hand is present, and falsely labels white surgical tissue and skin as hands. Our method, instead, can distinguish hands successfully from other structures in the scenes. Our method also missed fewer partial hands than other methods. We qualitatively observe in Figure 6, that our model detects only partially visible hands more often than the baseline methods. In Figure 7, we show failure cases of our model, including cases in which our method does not detect hands, misinterprets the shape of a hand, or mistakenly detects hands in surgical cloths.

4.3. Missing labels

To investigate how the unlabeled dataset is used during training, we make the task more difficult and use only a fraction of the labeled data according to the protocols used by Soft Teacher and STAC. The remaining labeled data is used as unlabeled data. Results are compared in Table 3. On all three data folds (1%, 5%, and 10%) all three semi-supervised methods outperform the unaugmented and augmented supervised baselines. Semi-supervised methods are thus especially useful when only very small amounts of labeled data are available. However, we expect a method to leverage unlabeled data in all data regimes, making our model more applicable to real world scenarios.

	1%	5%	10%
Supervised baseline	60.5±2.2	74.7±1.0	78.2±0.6
Supervised baseline w/ augment.	64.0±0.9	75.4±1.4	79.6±0.8
Soft Teacher	**68.2**±2.4	77.1±1.3	**80.4**±0.7
STAC	68.0±2.2	**78.0**±1.3	**80.4**±1.3
Ours	64.1±1.6	75.6±0.7	79.6±1.0

Table 3: Model comparison on AP@50 using the partially labeled data setting. Results are averaged over five different random subsets and standard deviation is reported. In the low-data regime semi-supervised methods clearly outperform the supervised baselines.

It is interesting to note that Soft Teacher is the best performing model on the 1% data fold, but among the worst performing models on the full dataset. We posit that this is due to Soft Teacher's approach of training the student and teacher model simultaneously, rather than using a multi-stage approach where the teacher model is fully trained prior to the student model. This means that pseudo labels produced by the teacher in the earlier iterations of training are a lot worse than the pseudo labels that would have been produced by a pretrained teacher model. When training on the full dataset, the gap between the pseudo label quality of early iterations of the Soft Teacher and a fully trained teacher is high. In the 1% case, performance of a fully pretrained teacher is lower, and thus the gap between the pseudo label quality of the early iterations of the Soft Teacher and a fully trained teacher model is lower, making Soft Teacher more comparable to STAC.

Access to professionally trained medical annotators can be limited, meaning that annotation quality of medical datasets can be a concern. A second experiment explores the effect of missing labels from the full labeled dataset. As the dataset is poorly labeled (see Figure 1), we are interested in how robust the methods are when some hands are not labeled in the labeled training set. To simulate missing labels, we randomly remove some individual labels (rather than entire images) from the dataset retaining only 70%, 50%, 30%, or 10% of labels during training. The results of this experiment are shown in Figure 8. Our method outperforms the supervised baseline, Soft Teacher, and STAC on all data folds. When most hands are missing, STAC and Soft Teacher fail to outperform the supervised baseline, although the models have access to unlabeled data during training. As in previous experiments, we hypothesize that the pseudolabel-based methods fail to learn usable pseudo labels from the low number of retained ground truth labels. Our consistency-based model, instead, is robust to missing labels.

4.4. Ablation study

We ablate our results on the test set, and investigate the effect of the IoU threshold and the top κ filtering in Tables 4 and 5. We use a high confidence threshold of 0.9 for this ablation study, to allow for direct comparability to STAC and our best performing model, but note that this reduces the effect of the exact IoU threshold value and top κ filtering on the performance.

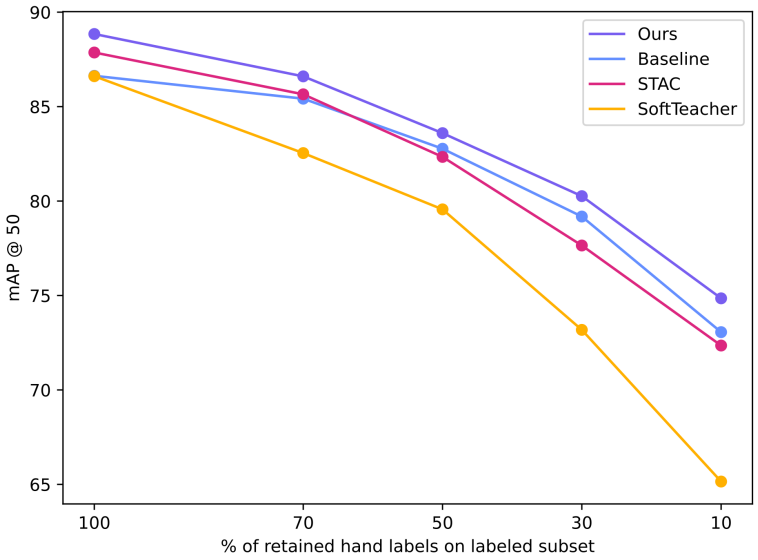

Figure 8: Model comparison on incompletely labeled data. Our method is more robust to missing labels in the ground truth than the semi-supervised baselines.

5. Discussion

The goal of this work is to detect surgeons' hands in videos. To this end, we compared three semi-supervised models and two supervised baselines, all built on the same backbone and differ mainly in their training strategies. Perhaps the most surprising insight is that the state-of-the-art method Soft Teacher does not outperform its supervised counterpart on the AVOS hands dataset. Adding unlabeled data decreases the accuracy of this model. This observation is in stark contrast to the original paper. As we use the authors' implementation, the only difference is the dataset. The heterogeneous nature of surgical data, with many different lighting settings, layouts, qualities, and objects, prevents the teacher network from learning useful pseudo labels. Instead, the network predicts pseudo labels that hurt the performance on the test set. These observations are reversed in the low-data setting. When using only 1% of the labels, the network learns to predict some pseudo labels that improve the performance of the student network. Although the overall mean precision is low, one can assume that the network learns some easy-to-learn hands, which it predicts in the unlabeled dataset, generating helpful pseudo labels.

STAC performs better than Soft Teacher on the AVOS dataset and even outperforms our network on low-data settings and AP @ 75 on the whole dataset. Yet, STAC is not end-to-end trainable and requires a teacher network to be trained first, making STAC harder to use than Soft Teacher or our end-to-end trainable proposed method. As the main difference between STAC and Soft Teacher is the fully pretrained teacher in STAC, we posit that the initial pretraining allows the network to generate more useful pseudo labels.

Our method outperforms all other methods when using 100% of the labeled data and all the available unlabeled data. Unlike STAC and Soft Teacher, our method does not depend

IoU threshold	5%	25%	50%	75%	90%	95%
AP @ 50	88.19	**88.84**	88.32	88.23	88.44	88.31

Table 4: Ablation: Effect of IoU threshold

Top κ	1	3	5	10
AP @ 50	87.96	88.34	**88.84**	88.53

Table 5: Ablation: Effect of number of hands retained

on pseudo labels. Incorrectly generated pseudo labels have a negative impact on the ability of a model to detect hands, and to learn from augmented versions of the unlabeled image. Our method encourages the model to produce equivalent predictions on unaugmented and augmented versions of an image regardless of the correctness of the prediction, forcing it to regularize itself and be more robust to making consistent predictions in spite of variations in the data often seen in this domain. Further, our method is compelling in a setting where data is poorly labeled, such as data from the internet or other alternative data sources. Our method ranks behind Soft Teacher and STAC in low data settings but still outperforms the supervised baselines. In summary, all three semi-supervised methods outperform the supervised baseline when only very few labeled images are available. When enough labeled data is available, careful augmentation in a supervised training paradigm can outperform semi-supervised methods on heterogeneous datasets.

Limitations In this paper we detect only one class. As Soft Teacher and STAC are designed for the COCO dataset including dozens of classes, the methods' full potential might not be reached. Further, we do not compare our method to one-stage models, as they are assumed to have lower accuracy. However, their inferiority on healthcare data is yet to be established. Additionally, we evaluate our method on one dataset only. Future work could generalize our method to different object classes in different datasets. Lastly, the AVOS dataset is poorly labeled. It is therefore difficult to correlate quantitative results with properties of the data.

6. Conclusions

This work establishes the first benchmark for hand detection during open surgery. We compare different state-of-the-art object detectors and find they do not generalize well to diverse surgical videos. While strong in low-data settings, these existing methods can be inferior to supervised baselines when enough labeled data is available. Furthermore, we introduce a novel semi-supervised object detector for hands that does not depend on pseudo labels and instead employs a localization consistency loss. While this loss was previously only applicable to one-stage detectors, we generalized it to a two-stage detector incorporating domain knowledge. Our method improves supervised baselines in diverse data availability settings and is more robust towards diverse and poorly labeled data than previous semi-supervised methods while being end-to-end trainable. We demonstrate that unlabeled data scraped from the internet can improve hand detection results, making it a valuable alternative to labeled data, which is hard to collect in healthcare settings.

References

Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 06 2010. doi: 10.1007/s11263-009-0275-4.

Emmett D. Goodman, Krishna K. Patel, Yilun Zhang, William Locke, Chris J. Kennedy, Rohan Mehrotra, Stephen Ren, Melody Y. Guan, Maren Downing, Hao Wei Chen, Jevin Z. Clark, Gabriel A. Brat, and Serena Yeung. A real-time spatiotemporal AI model analyzes skill in open surgical videos. *CoRR*, abs/2112.07219, 2021. URL https://arxiv.org/abs/2112.07219.

Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/d0f4dae80c3d0277922f8371d5827292-Paper.pdf.

Wenjing Jiang, Tong Xia, Zhiqiong Wang, and Fucang Jia. Semi-supervised surgical tool detection based on highly confident pseudo labeling and strong augmentation driven consistency. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings*, page 154–162, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-88209-9. doi: 10.1007/978-3-030-88210-5_14. URL https://doi.org/10.1007/978-3-030-88210-5_14.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

Nathan Louis, Luowei Zhou, Steven J Yule, Roger D Dias, Milisa Manojlovich, Francis D Pagani, Donald S Likosky, and Jason J Corso. Temporally guided articulated hand pose tracking in surgical videos. *International Journal of Computer Assisted Radiology and Surgery*, 18(1):117–125, 2023.

Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *CoRR*, abs/2005.04757, 2020. URL https://arxiv.org/abs/2005.04757.

Farhana Sultana, Abu Sufian, and Paramartha Dutta. A review of object detection models based on convolutional neural network. *Intelligent computing: image processing based applications*, pages 1–16, 2020.

Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2291–2301, 2021.

Mansoor Ali Teevno, Gilberto Ochoa-Ruiz, and Sharib Ali. A semi-supervised teacher-student framework for surgical tool detection and localization. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 0(0):1–9, 2022. doi: 10.1080/21681163.2022.2150688. URL https://doi.org/10.1080/21681163.2022.2150688.

Jinman Wei, Jinkun Yao, Guoshan Zhanga, Bin Guan, Yueming Zhang, and Shaoquan Wang. Semi-supervised object detection based on single-stage detector for thighbone fracture localization, 2022.

Yue Wu, Yang Zhou, Jianchun Zhao, Jingyuan Yang, Weihong Yu, Youxin Chen, and Xirong Li. Lesion localization in OCT by semi-supervised object detection. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*. ACM, jun 2022. doi: 10.1145/3512527.3531418. URL https://doi.org/10.1145%2F3512527.3531418.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021.

Jihun Yoon, Jiwon Lee, SungHyun Park, Woo Jin Hyung, and Min-Kook Choi. Semi-supervised learning for instrument detection with a class imbalanced dataset. In Jaime Cardoso, Hien Van Nguyen, Nicholas Heller, Pedro Henriques Abreu, Ivana Isgum, Wilson Silva, Ricardo Cruz, Jose Pereira Amorim, Vishal Patel, Badri Roysam, Kevin Zhou, Steve Jiang, Ngan Le, Khoa Luu, Raphael Sznitman, Veronika Cheplygina, Diana Mateus, Emanuele Trucco, and Samaneh Abbasi, editors, *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 266–276, Cham, 2020. Springer International Publishing. ISBN 978-3-030-61166-8.

Michael Zhang, Xiaotian Cheng, Daniel Copeland, Arjun Desai, Melody Y Guan, Gabriel A Brat, and Serena Yeung. Using computer vision to automate hand detection and tracking of surgeon movements in videos of open surgery. In *AMIA Annual symposium proceedings*, volume 2020, page 1373. American Medical Informatics Association, 2020.

Hong-Yu Zhou, Chengdi Wang, Haofeng Li, Gang Wang, Shu Zhang, Weimin Li, and Yizhou Yu. Ssmd: Semi-supervised medical image detection with adaptive consistency and heterogeneous perturbation, 2021.