# A semi-supervised Teacher-Student framework for surgical tool detection and localization

Mansoor Ali Teevno[a], Gilberto Ochoa-Ruiz[a], and Sharib Ali[b]

[a]Tecnologico de Monterrey, Escuela de Ingeniería y Ciencias, 38115, México; [b]School of Computing, University of Leeds, Leeds, UK

**ARTICLE HISTORY**
Compiled September 4, 2022

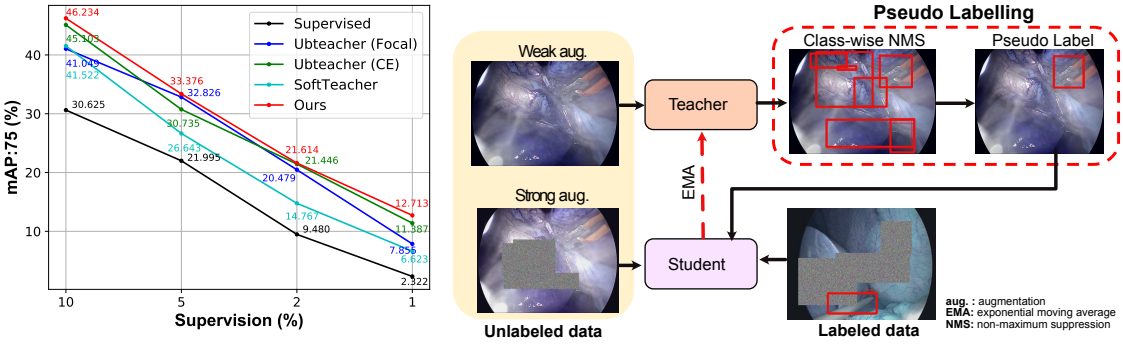
**ABSTRACT**
Surgical tool detection in minimally invasive surgery is an essential part of computer-assisted interventions. Current approaches are mostly based on supervised methods which require large fully labeled data to train supervised models and suffer from pseudo label bias because of class imbalance issues. However large image datasets with bounding box annotations are often scarcely available. Semi-supervised learning (SSL) has recently emerged as a means for training large models using only a modest amount of annotated data; apart from reducing the annotation cost. SSL has also shown promise to produce models that are more robust and generalizable. Therefore, in this paper we introduce a semi-supervised learning (SSL) framework in surgical tool detection paradigm which aims to mitigate the scarcity of training data and the data imbalance through a knowledge distillation approach. In the proposed work, we train a model with labeled data which initialises the Teacher-Student joint learning, where the Student is trained on Teacher-generated pseudo labels from unlabeled data. We propose a multi-class distance with a margin based classification loss function in the region-of-interest head of the detector to effectively segregate foreground classes from background region. Our results on m2cai16-tool-locations dataset indicate the superiority of our approach on different supervised data settings (1%, 2%, 5%, 10% of annotated data) where our model achieves overall improvements of 8%, 12% and 27% in mAP (on 1% labeled data) over the state-of-the-art SSL methods and a fully supervised baseline, respectively. The code is available at https://github.com/Mansoor-at/Semi-supervised-surgical-tool-detection.

**KEYWORDS**
Semi-supervised learning, Faster-RCNN, Surgical tool detection

## 1. Introduction

Recent works in deep learning based on visual recognition methods have delivered enormous advantages towards computer-assisted interventions (CAI) (Ward et al. 2021). CAI tools have been primarily focused on specific information gathering, such as the presence or location of lesions. Nonetheless, recent developments in the image recognition tasks with improved accuracy have led to expansion of its scope to several other areas including intra-operative decision support systems (Bouget et al. 2017). These applications provide contextual information to the surgeon during the surgery, as a post-operative feedback (Bhatia et al. 2007; Sarikaya et al. 2017) for surgical training and video content analysis (Wang et al. 2019).

CONTACT Sharib Ali. Email: s.s.ali@leeds.ac.uk




(a) (b)

Figure 1.: **Comparison at different percentage of supervision and augmentation strategies in a Teacher-Student paradigm.** (a) The proposed approach efficiently leverages unlabeled data and produces substantial improvement over supervised baseline and on the SOTA Unbiased Teacher (Ubteacher) Liu et al. (2021) framework. (b) Data augmentation workflow of the proposed Teacher-Student mutual learning approach. Unlabeled data given with weak augmentation to Teacher and with strong augmentation to Student. Student gets pseudo labels through NMS and thresholding from the Teacher.

More recently, CAI systems capable of performing effectively the sub-tasks such as surgical phase recognition, identification of the tool presence, as well as their recognition, localization and instance-based segmentation are getting increased attention (Bouget et al. 2017). The development of these task-based automated approaches can ensure improved surgical care, patient safety and alleviate surgeon fatigue.

Deep learning based surgical tool detection task has attracted a lot of attention in recent years. However, most of the state-of-the-art (SOTA) methods have employed fully supervised approaches (Jin et al. 2018; Zhang et al. 2020) and only a few weakly supervised methods, mostly implementing classification models for determining tool presence (Vardazaryan et al. 2018) have been proposed. Nonetheless, training complex deep learning (DL) models under the supervised setting requires difficult-to-acquire and precisely annotated datasets, which is a time consuming task and susceptible to intra and inter-observer bias in annotations. As a result, only a few labeled surgical tool datasets are publicly available (Sarikaya et al. 2017; Jin et al. 2018) and this lack of annotated datasets has essentially hindered the development of robust and generalizable deep architectures for the surgical instrument detection.

Alternatively, the annotation cost could be greatly mitigated by exploiting unlabeled data through efficient semi-supervised learning (SSL) frameworks. The core idea of SSL is to be able to extract information from the unlabeled data that is essential for label prediction. One solution is to train a network to solve a pre-defined pretext task (Teacher model generating pseudo labels) and then using the learned knowledge in the downstream task (Student network). Recently, SSL has shown promising outcomes in improving model performance and is receiving growing attention of the computer vision research community (Van Engelen and Hoos 2020; Sohn et al. 2020). Despite these progresses, most of these advances are in the domain of image classification rather than object detection as the bounding box annotations require more time and effort to generate. Traditionally, SSL can be approached with adapting SOTA image classification methods such as (Sohn et al. 2020) to object detection. However, existence of some unique characteristics such as foreground-background and foreground class imbalance makes object detection interact poorly with those methods. The class imbal-

ance problem may greatly impede the use of pseudo-labeling based training pipelines since Teacher generated pseudo labels will be overly biased towards dominant classes and ignoring minor and less dominant classes. As a result, these models in their vanila arrangement will exacerbate the class-imbalance problem and cause severe overfitting. To overcome these issues, we propose a jointly trained Teacher-Student model on m2cai16-tool-locations dataset (Jin et al. 2018) which is initialised by a supervised detector. We argue that slowly updating the Teacher by exponential moving average (EMA) via the Student can alleviate pseudo-labeling bias problem and improve pseudo label quality, hence overall performance improvement. Additionally, we propose a multi-class distance and margin-based classification loss in the ROI head of the detector network to boost the classification performance. This is achieved by maximising the distance between foreground classes and the background. To the best of our knowledge, our approach is the first effort towards leveraging Teacher-Student joint training paradigm for addressing data scarcity problem in surgical tool detection applications. We employ strong and weak augmentation pipelines to improve model robustness (Fig. 1(b)). Our proposed pipeline outperforms supervised baseline and other SOTA semi-supervised methods in terms of classification and localisation performance (Fig. 1(a)).

In the rest of the paper, we discuss related work (Section 2), materials and method (Section 3), quantitative and qualitative results (Section 5), ablation studies (Section 5.3) and lastly discussion and conclusion (Section 6).

## 2. Related Work

Some of the early works on surgical tool detection used radio frequency identification tags (Kranzfelder et al. 2013), Viola-Jones detection algorithm (Lalys et al. 2011) and segmentation, contour delineation and three-dimensional modeling (Speidel et al. 2009). With the advent of deep learning-based approaches using convolutional neural networks, computer vision methods have evolved with remarkable growth and demonstrated promising outcomes (Russakovsky et al. 2015). In the surgical domain, several works have leveraged deep learning approaches to obtain SOTA performance on surgical instrument detection (Jin et al. 2018; Sahu et al. 2016; Twinanda et al. 2016a,b). Most of the studies conducted on surgical tool detection have proposed supervised pipelines or only have targeted frame-level tool presence detection. For example, AG-Net (Hu et al. 2017) used global and local prediction networks to obtain visual cues for tool presence detection and showed a significant improvement over m2cai16-tool challenge (Raju et al. 2016) winners. Jin *et al.* (Jin et al. 2018) proposed region-based convolutional neural network to perform surgical skill assessment adapted to tool presence detection, spatial localization and tracking. The authors also extended the m2cai16-tool dataset (dat 2016) to include tool bounding boxes (subsequently named as m2cai16-tool-locations) which we have used in this work. Sarikaya *et al.* used image and temporal motion cues to train multi-modal CNN models (Sarikaya et al. 2017) for tool detection and localization in robotic-assisted surgical training task videos. Tool detection and pose estimation was also studied in (Reiter et al. 2012) but it was limited to robotic arms that return kinematic data. Shi *et al.* proposed a lightweight attention-guided framework (Shi et al. 2020a) for tool detection and conducted an ablation study on three different datasets (two public datasets, EndoVis Challenge (Kurmann et al. 2017) and ATLAS Dione (Sarikaya et al. 2017) and one self-prepared cholec80-locations). However, their model performed well on all tools except grasper and irrigator classes. In another study (Zhang et al. 2020), irrigator can be observed as worst performing instrument with average precision of 41.6%, followed by grasper with

54.1% in a supervised setting at IOU threshold of 50%. A ghost feature maps-based pipeline was used to reduce the computational burden for tool detection in (Yang et al. 2021). A CNN-based hidden Markov model was proposed by Twinanda in Twinanda et al. (2016a) for surgical tool detection from laparoscopic videos. A combination of CNN to extract spatial features and long short-term memory (LSTM) for temporal cues was proposed to perform surgical tool detection from laparoscopic videos (Mishra et al. 2017).

Although the results of some of these approaches have been mostly encouraging, they have reported only one mAP results (Sarikaya et al. 2017; Shi et al. 2020b) which is not quite sufficient to gauge the classification and localisation performance. Furthermore, previous approaches require completely labeled datasets to train the model. Such datasets are either scarcely available or the process of annotating them can lead to other issues such as introducing unintended biases in the trained model.

In this work, we aim to demonstrate the advantages of an SSL approach and propose a novel semi-supervised Teacher-Student framework to alleviate the limited data problem and annotation cost requirement for training on larger datasets. Our literature search revealed that there are only two studies conducted on semi-supervised learning in the medical domain where one is based on cataract surgery dataset (Jiang et al. 2021) while another study (Yoon et al. 2020) used a tracker to detect instruments from unlabeled private surgery videos. To the best of our knowledge, this is the first approach that investigates the effectiveness of unlabeled data through a Teacher-Student learning pipeline for tool detection on a minimally invasive surgery dataset. We report results from our model in terms of mAP on various IOU thresholds to demonstrate the effectiveness of our approach in detecting and localising surgical tools.

## 3. Materials and Method

### *3.1. Dataset*

In this work, we use an extended version of the m2cai16-tool dataset which was originally released for M2CAI 2016 Tool Presence Detection Challenge (Twinanda et al. 2016b). This dataset consists of 15 videos each with duration from 20 to 75 minutes of cholecystectomy procedures performed at the University Hospital of Strasbourg in France. After down sampling at 1 fps, it leaves 23,000 frames annotated with tool presence classification.

Later, m2cai16-tool-locations dataset was build with spatial bounding box annotations (Jin et al. 2018). This dataset consists of a total of 2812 frames that were annotated under supervision and spot-checking from clinical experts. We have used 80%, 10%, and 10% for training, validation and test splits, respectively. The annotations breakdown per class is given in **supplementary material (Table 1)** and the tool instances with example box annotations are presented in Fig. 2.

We use average precision (AP) computed per class and mean average precision (mAP) for all seven classes which are the standard object detection evaluation metric. These metrics are evaluated at different IoU thresholds, usually denoted as $mAP_{IoU-threshold}$. We report results for 50, 75, 50:95 (average of AP values for IoU thresholds from 50 to 95 with interval of 5), medium and large IoU thresholds.

### *3.2. Data Augmentation*

We have used two data augmentation strategies in this work, which we refer as weak and strong augmentations (Fig. 1(b)). For the weak augmentation, we apply random horizontal flips whilst for strong augmentation, we randomly perform several photometric augmentations like grayscale, color jittering, Gaussian blur, patch masking

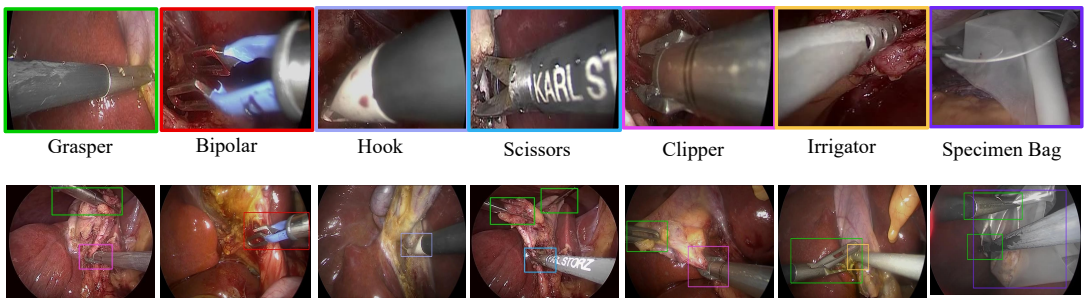



Figure 2.: (Top) Samples from *m2cai16-tool-locations* dataset with representative classes of seven tools. (Bottom) Example frames with bounding box annotations where color of the box refers to tool class

and cutout patches DeVries and Taylor (2017). For the complete description of data augmentation with parameter values, please refer to Liu et al. (2021).

## 4. Method

In this work we address multi-instance surgical tool detection problem in a semi-supervised setting. Let the training set in various arrangements of labeled data sets be denoted as $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and unlabeled data sets be $D_u = \{x_i^u\}_{i=1}^{N_u}$, where $N_s$ and $N_u$ represent number of supervised and unsupervised training samples while $y^s$ represent bounding box annotation of each labeled image $x^s$. Here, $y^s$ consists of bounding boxes for all object instances, height and width of image and instance category names. It is important to mention that since all the training data samples contain labels, during training we removed the labels of the portion we categorise as unlabeled. The overall training pipeline is divided into two stages as shown in Fig. 3. The first stage is the initialization stage (section 4.1), while the second is the Teacher-Student joint learning mechanism (section 4.2). In the second stage, the Teacher generates pseudo-labels and Student network is trained on both pseudo labeled data and supervised data. Each stage is detailed separately below along with Student learning and Teacher update scheme and margin .

### *4.1. Initialization stage*

The initialization stage acts as a trigger point for Teacher-Student joint learning. It sets the stage for the Teacher model to be able to generate qualitative pseudo-labels for better Student learning. In this stage, we exploit the available labeled data $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ to train the Faster-RCNN detector model ($\theta$) with supervised loss $\mathcal{L}_{sup}$. The standard Faster-RCNN model makes use of four losses: RPN classification loss $\mathcal{L}_{cls}^{rpn}$, RPN regression loss $\mathcal{L}_{reg}^{rpn}$, ROI classification loss $\mathcal{L}_{cls}^{roi}$ and ROI regression loss $\mathcal{L}_{reg}^{roi}$ (Eq. (1)).

$$\mathcal{L}_{sup} = \sum_{i}^{N_s} \mathcal{L}_{cls}^{rpn}(x_i^s, y_i^s) + \mathcal{L}_{reg}^{rpn}(x_i^s, y_i^s) + \mathcal{L}_{cls}^{roi}(x_i^s, y_i^s) + \mathcal{L}_{reg}^{roi}(x_i^s, y_i^s) \tag{1}$$

The weights and architecture of the model trained during this initialization phase are then copied to be used for both the Student and Teacher models ($\theta_{\mathcal{T}} \leftarrow \theta, \theta_{\mathcal{S}} \leftarrow \theta$). The trained detector from this stage provides a good initialization for next stage, where we further exploit unsupervised data to improve object detection.

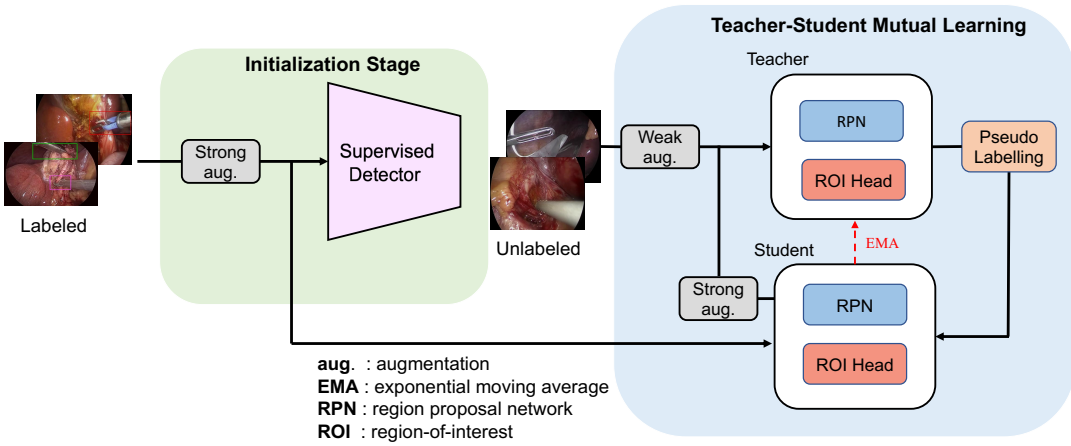



Figure 3.: Overview of the proposed Surgical tool detection model. It consists of two modules: 1) An initialisation module, where a supervised model makes use of strongly augmented labeled data, 2) A Teacher-Student mutual learning module, where the Student is trained with strongly augmented unlabaled data with Teacher-generated pseudo labels. The Student transfers learned weights to the Teacher gradually through Exponential Moving Average (EMA).

### *4.2. Teacher-Student joint learning stage*

The proposed knowledge distillation framework leverages Student and Teacher joint training to address lack of data problem. During training, Teacher generates pseudo labels on unlabeled data and Student is trained on those labels. Thus, a continuously learning Student passes on the learned knowledge to the Teacher. We posit that this evolving mutual learning would result in better detection performance by generating stable and reliable pseudo labels. Weak and strong augmentation pipelines ensure reliable pseudo label generation by Teacher and diversity in Student models respectively.

### *4.3. Student learning and Teacher update scheme*

We tackle the pseudo-label noise problem which may cause severe performance degradation (Sohn et al. 2020) by confidence thresholding ($\tau$). Although this step could have sufficed in the case of image classification, for object detection tasks, additional steps must be enforced as duplicated bounding box predictions and class imbalanced prediction problems are typically encountered in these settings. We address the duplicated box predictions problem by applying class-wise non-maximum suppression (NMS) before a confidence thresholding step. As simple confidence thresholding only removes samples with low confidence on predicted object categories and does not take into account the quality of bounding box locations, we do not use unsupervised loss on bounding box regression which is thus represented as below with $\theta_S$ as weight updates between both supervised $\mathcal{L}_{sup}$ and unsupervised $\mathcal{L}_{unsup}$ losses:

$$\mathcal{L}_{unsup} = \sum_{i}^{N_u} \mathcal{L}_{cls}^{rpn}(x_i^u, \tilde{y}_i^u) + \mathcal{L}_{cls}^{roi}(x_i^u, \tilde{y}_i^u) \tag{2}$$

$$\theta_{\mathcal{S}} \leftarrow \theta_{\mathcal{S}} + \gamma \frac{\partial(\mathcal{L}_{sup} + \lambda_u \mathcal{L}_{unsup})}{\partial \theta_{\mathcal{S}}}, \tag{3}$$

where $\gamma$ is the learning rate and $\lambda_u$ is unsupervised loss weight. The overall unsupervised loss in Eq. (2) consists of the sum of RPN and ROI head classification losses.

Eq. (3) depicts the Student weight update scheme which includes both supervised and unsupervised losses with a loss weight parameter $\lambda_u$.
Finally, we perform Teacher model refinement by using EMA following *Mean Teacher* to slowly update Teacher network which in turn will generate stable and reliable pseudo labels. The update can be represented as:

$$\theta_{\mathcal{T}} \leftarrow \alpha\theta_{\mathcal{T}} + (1-\alpha)\theta_{\mathcal{S}}, \tag{4}$$

where $\alpha$ is the EMA rate, and $\theta_{\mathcal{T}}$, $\theta_{\mathcal{S}}$ are the network weights for Teacher and Student.

### *4.4. Logistic loss with added margin and distance penalization*

In the surgical domain, foreground class imbalance exists in every dataset due to the fact that tool usage frequency varies from one tool to another (Mishra et al. 2017). In this work, to address the class imbalance problem, we target the foreground and background class imbalance problem by introducing a multi-class loss function based on a margin, which tries to maximise foreground-background distance. Unlike the focal or cross entropy losses, our proposed loss tries to predict relative distance between inputs. Specifically, we divide classification logits between foreground and background instances and then compute *sigmoid* probability, respectively. We then sum the *softmax* of the probabilities over all the batch for the foreground $\rho$ and background $\beta$ instances. These probabilities are then used to maximise foreground-background distance in the final loss computation which is in the form of a logistic loss function for classification defined as:

$$\mathcal{L}_{cls}^{roi} = \sum_{n} w_l \log(1 + \frac{e^{s \cdot (\beta - \rho + \sigma)}}{s}), \tag{5}$$

where $n$ is the mini-batch size, $w_l$ represents loss weight, s is the smoothness parameter and $\sigma$ denotes margin.
Apart from the multi-class loss, Teacher update with EMA will also help reduce pseudo label bias since new Teacher is regularised by previous Teacher model which prevents drastic movement of the decision boundary towards under-represented classes.

**Algorithm 1** Multi-class distance and margin based classification loss

```
procedure LOSS(logits, targets)
    classes ← class_indices
    fg_logits ← logits(targets = classes)
    bg_logits ← logits(targets! = classes)
    fg_prob ← sigmoid(fg_logits)
    bg_prob ← sigmoid(bg_logits)
    ρ ← ∑ softmax(fg_prob)
    β ← ∑ softmax(bg_prob)
    loss ← Eq : 5
end procedure
```

## 5. Experiments and results

### *5.1. Implementation Details*

The implementation of our proposed framework is based on Faster-RCNN detector model with ResNet50-FPN backbone, whose network weights are initialized by ImageNet pretrained model. We use a confidence threshold ($\tau$) of 0.7, regularization co-efficient for unsupervised loss ($\lambda_u$) of 0.2 and EMA rate ($\alpha$) of 0.9996. We use *WarmupMultiStepLR* as a learning rate ($\alpha$) scheduler in initialization stage while a constant learning rate of 0.01 for the Teacher-Student mutual learning stage. In the initialization stage, we use strong augmentation, while during the Teacher-Student mutual learning, we use weak augmentation for the Teacher and strong augmentation for Student. We report results in terms of mAP on different IOU thresholds. We use a batch size of 8 (4 labeled images and 4 unlabeled images) throughout the experiments. We performed network training through detectron2 (Wu et al. 2019) object detection framework using 4 GPUs on NVIDIA Tesla P100-SXM2-16GB system. We use fixed seed values for generating the data splits to make the results more reproducible.

### *5.2. Results*

#### *5.2.1. Quantitative Results*

We evaluate our model with different labeled and unlabeled data protocols and present the results on a 10% held-out set in Table 1. The table also includes results on the supervised baseline, UnbiasedTeacher (Liu et al. 2021) with both CrossEntroppy and focal losses. and SoftTeacher (Xu et al. 2021). Table 2 shows per class $mAP_{50:95}$ results on 1% labeled data setting. Furthermore, we also conduct a paired t-test between $AP_{50}$ obtained by our proposed model and $AP_{50}$ obtained by other SOTA methods. The resulting box-plot on 1%, 2%, 5% and 10% labeled data setting are shown in Fig.5 and $p$-values are shown in Table 1.

#### *5.2.2. Qualitative Results*

In this section, we report the qualitative performance of our model as shown in Fig. 4. The example surgical scenes are carefully chosen to contain several instances in one frame (column two from left), only partially visible instrument (column three from left), irregular orientation (column four from left). Results on all data settings have been presented to see how well model performs in terms of detection and localisation.

### *5.3. Ablation Study*

Several ablation studies were conducted to validate the effectiveness of different parameters. We evaluated the effect of initialization, confidence threshold ($\tau$), EMA rates and normalization parameter (s) on model performance. We trained the model with and without initialization stage and concluded that such process does improve the overall performance by a substantial margin (**Supplementary material section 1.1**). We also evaluated the model on different values of $\tau$ where $\tau$=0.7 gives the best performance (**Supplementary material section 1.2**). We also performed multiple experiments to evaluate the impact of Teacher update with EMA rate on model performance for which EMA of 0.9996 gave the optimum performance (**Supplementary material section 1.3**).
Here, we present ablation for use of different loss functions and our proposed loss with different normalization parameter 's' values. It can be observed that our proposed loss with $s = 5$ provided the best performance with the highest mAP over all IoU thresholds (see Table 3).

Table 1.: Experimental results with ResNet50-FPN as backbone .

| Method | $\text{mAP}_{50}$ | $\text{mAP}_{50:95}$ | $\text{mAP}_{75}$ | $\text{mAP}_{m}$ | $\text{mAP}_{l}$ | $p$-values |
|---|---|---|---|---|---|---|
| | **1% Labeled data** | | | | | |
| Supervised | 23.578 | 7.673 | 2.322 | 6.189 | 9.050 | 5.996e-17 |
| Unbiased Teacher* (Liu et al. 2021) | 34.374 | 14.145 | 7.855 | 10.687 | 15.880 | 5.626e-02 |
| Unbiased Teacher** (Liu et al. 2021) | 42.382 | 18.008 | 11.387 | 13.041 | 20.135 | 6.229e-03 |
| SoftTeacher (Xu et al. 2021) | 38.421 | 13.556 | 6.623 | 16.756 | 13.045 | 5.526e-02 |
| Ours | 50.632 | 20.094 | 12.713 | 15.219 | 21.774 | – |
| | **2% Labeled data** | | | | | |
| Supervised | 47.140 | 18.609 | 9.480 | 24.033 | 18.586 | 2.558e-14 |
| Unbiased Teacher* (Liu et al. 2021) | 71.608 | 31.752 | 20.479 | 27.871 | 32.430 | 3.975e-04 |
| Unbiased Teacher** (Liu et al. 2021) | 72.416 | 31.490 | 21.446 | 26.767 | 32.666 | 2.010e-01 |
| SoftTeacher (Xu et al. 2021) | 60.366 | 25.421 | 14.767 | 17.991 | 28.323 | 2.558e-8 |
| Ours | 72.341 | 32.311 | 21.614 | 29.780 | 33.556 | – |
| | **5% Labeled data** | | | | | |
| Supervised | 71.082 | 32.249 | 21.995 | 29.505 | 35.041 | 5.866e-05 |
| Unbiased Teacher* (Liu et al. 2021) | 84.721 | 42.269 | 32.826 | 35.697 | 44.204 | 1.298e-01 |
| Unbiased Teacher** (Liu et al. 2021) | 82.592 | 40.393 | 30.735 | 33.665 | 42.904 | 3.424e-04 |
| SoftTeacher (Xu et al. 2021) | 83.211 | 38.857 | 26.643 | 30.567 | 40.718 | 4.566e-04 |
| Ours | 84.427 | 42.392 | 33.376 | 31.156 | 44.610 | – |
| | **10% Labeled data** | | | | | |
| Supervised | 80.193 | 38.640 | 30.625 | 29.845 | 40.958 | 7.108e-03 |
| Unbiased Teacher* (Liu et al. 2021) | 92.981 | 47.369 | 41.049 | 41.137 | 48.714 | 2.291e-01 |
| Unbiased Teacher** (Liu et al. 2021) | 90.353 | 45.972 | 45.103 | 39.247 | 47.787 | 1.00 |
| SoftTeacher (Xu et al. 2021) | 89.362 | 42.717 | 41.522 | 38.312 | 43.849 | 1.00 |
| Ours | 90.250 | 46.886 | 46.234 | 42.635 | 48.644 | – |

* with Focal Loss ; ** with Cross Entropy Loss.

Table 2.: The average precision ($AP_{50:95}$) per class on 1% labeled data.

| Class | Supervised | Ubteacher* | Ubteacher** | SoftTeacher | Ours |
|---|---|---|---|---|---|
| Grasper | 12.434 | 23.457 | 23.046 | 7.0 | 20.203 |
| Bipolar | 13.450 | 19.472 | 33.384 | 22.2 | 29.499 |
| Hook | 11.349 | 38.529 | 44.614 | 29.8 | 43.924 |
| Scissors | 3.592 | 4.130 | 5.052 | 3.8 | 6.860 |
| Clipper | 4.273 | 5.045 | 4.800 | 0.0 | 4.970 |
| Irrigator | 4.022 | 3.393 | 8.468 | 27.9 | 10.331 |
| SpecimenBag | 4.592 | 4.986 | 6.692 | 4.1 | 24.873 |

* Unbiased Teacher with focal loss; ** Unbiased Teacher with cross entropy loss;

## 6. Discussion and conclusion

We demonstrate that our proposed approach performs favourably against the SOTA semi-supervised models (Liu et al. 2021) and (Xu et al. 2021) . In 1% setting our proposed model outperforms unbiased Teacher with focal loss by a large margin and cross entropy loss by a 8 points on every evaluation metric while also outperforming SoftTeacher (Xu et al. 2021) model. It is worth noting that our approach achieves 50.632% $mAP_{50}$ on 1% labeled data which is even higher than supervised baseline trained on 2% labeled data and this trend can be witnessed in all settings. This improvement can be attributed to several crucial factors such as gradual improvement in pseudo label

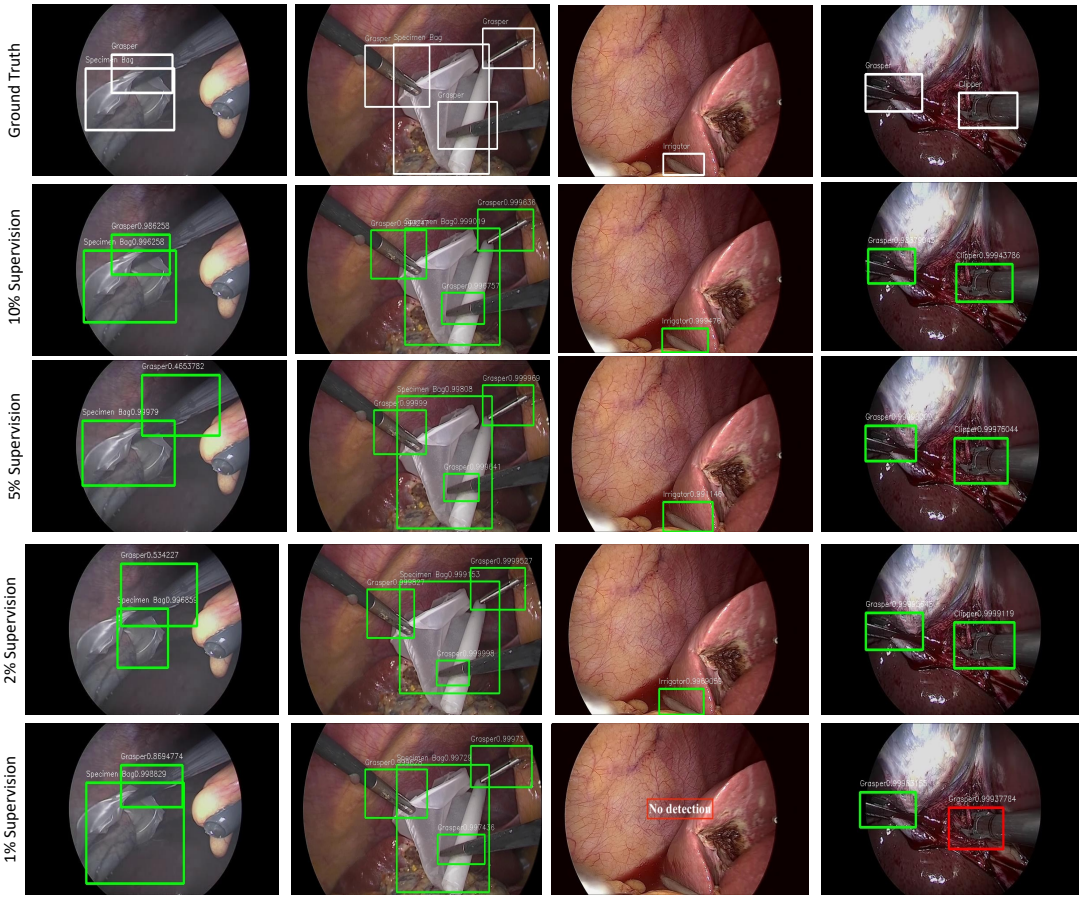



Figure 4.: Qualitative Results: First row shows images with ground truth. Second, third, fourth and fifth row presents results on 10%, 5%, 2% and 1% setting respectively, Green and red boxes indicate correct and wrong predictions

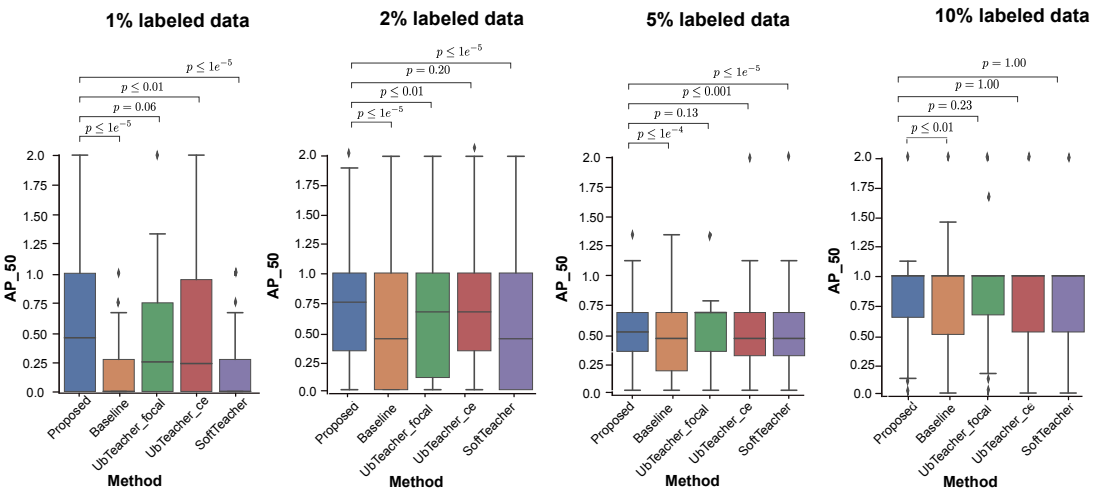



Figure 5.: Box-plots for paired t-test on 1%, 2%, 5% and 10% labeled data setting. Here the AP_50 scores above 1 are only shown to represent standard deviation in the scores.

quality through EMA training which is in contrast to previous approaches in which Teacher model is freezed after training on labeled data. Another factor is the intro-

Table 3.: Normalization parameter $s$ grid search

| Loss | $mAP_{50:95}$ | $mAP_{50}$ | $mAP_{75}$ | $mAP_m$ | $mAP_l$ |
|---|---|---|---|---|---|
| Focal | 14.145 | 34.374 | 7.855 | 10.687 | 15.880 |
| Cross Entropy | 18.008 | 42.382 | 11.387 | 13.041 | 20.135 |
| Proposed loss (s=3) | 16.260 | 41.438 | 8.260 | 10.801 | 18.848 |
| Proposed loss (s=4) | 18.475 | 44.534 | 9.252 | 13.246 | 20.483 |
| Proposed loss (s=5) | **20.094** | **50.632** | **12.713** | **15.216** | **21.774** |
| Proposed loss (s=6) | 18.993 | 47.597 | 11.156 | 10.380 | 22.006 |

duction of loss function which effectively increases the foreground-background distance and helps in improving detection performance.

Furthermore, the proposed framework performs much better on $mAP_{75}$ in all settings consistently which indicates improved localisation performance. On the 2% labeled data setting, our model obtained 72.341% $mAP$on 50% IoU thresholds, while unbiased Teacher on focal and cross entropy losses achieved 71.608% and 72.416% which is just slightly greater than our model. However, if we compare the performance of our model on $mAP$ at 50:95 and 75 IoU thresholds, we observe that our model consistently gives superior performance. Moreover, on 5% and 10% setting, Unbiased Teacher (Liu et al. 2021) with focal loss achieves slightly higher performance on $mAP_{50}$, but the proposed method gives superior performance in terms of $mAP_{50:95}$ and $mAP_{75}$. This validates the effectiveness of our method on both classification and localization performance.

We also present per class average precision ($AP_{50:95}$) in Table 2. Here we observe significant improvement in average precision on all instances specially hard-to-detect classes like Specimen Bag (20 points), Irrigator (6 points) and Bipolar (16 points) against the supervised baseline.

The qualitative results also indicate strong performance of our approach as most of the tools (even when four tools in one frame) are detected and localised correctly. The localisation accuracy increases as we add more labeled data as is evident from Fig. 4 from bottom to top, however the detection performance remains largely unchanged. There are some missed detections on 1% of the labeled data setting (see row 5 in column 3) and incorrect class label prediction (see row 5 in column 4). The missed detection occurred mostly on 1% labeled data setting where model did not see enough annotated examples. Incorrect class prediction in the bottom right may be due to less discrminative features between both instances. Similarly, the missed detection in second last image on the bottom row can be because the tool was only partly visible.

The paired t-test $p$-values computed between the proposed method and SOTA methods are given in the Table 1. Also, we have shown a bar-plot with median and deviations and significance between the SOTA and proposed methods (see Fig. 5). We can observe that our proposed approach performs well on different data settings. From Fig. 5, it can be observed that for 1% setting our method is significantly different compared to other SOTA methods with the highest median $AP_{50}$ value reported. Similarly, on the 2% setting our model and Unbiased Teacher model on cross entropy loss (UbTeacher_ce) performed equally well ($p$-value = 0.20) but still with the highest median value compared to other methods. Similar performance changes can be observed for 5% data where Unbiased Teacher model on focal loss (UbTeacher_focal) has $p$-value = 0.13 (computed at $AP_{50}$) while on $mAP_{75}$ our method is still better. The reason behind competitive scores in these cases is because the reported APs are only done

at 50% IoU threshold, while it is evident from Table 1 that our method performance for other mAPs at higher IoU thresholds has distinguishable improvements. However, with the 10% labeled data setting, we reach non-significant difference in $p$-values for other unbiased models and SoftTeacher model. This is because 10% in this case is enough data for supervision during training.

In this paper, we addressed a lack of annotated data problem in surgical domain for the first time by proposing a knowledge distillation framework. We tackle a multi-label, multi-class detection problem by implementing an end to end Teacher-Student learning with a multi-class foreground-background distance loss. We used strong and weak augmentation strategies to improve model robustness and class-wise NMS and EMA to improve pseudo label quality. Our experiments on m2cai16-tool dataset show the effectiveness of our model in terms of mAP on various supervision protocols against SOTA semi-supervised models. We also conducted extensive ablation experiments to demonstrate the validity of our proposed framework.

**Acknowledgments**

The authors wish to thank the AI Hub and the CIIOT at ITESM for their support for carrying the experiments reported in this paper in their NVIDIA's DGX computer.

**References**

2016. Tool presence detection challenge results. `http://cammau-strasbgfr/m2cai2016/indexphp/tool-presence-detection-challenge-results/`.

Bhatia B, Oates T, Xiao Y, Hu P. 2007. Real-time identification of operating room state from video. In: AAAI; vol. 2. p. 1761–1766.

Bouget D, Allan M, Stoyanov D, Jannin P. 2017. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. Medical image analysis. 35:633–654.

DeVries T, Taylor GW. 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:170804552.

Hu X, Yu L, Chen H, Qin J, Heng PA. 2017. Agnet: attention-guided network for surgical tool presence detection. In: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; p. 186–194.

Jiang W, Xia T, Wang Z, Jia F. 2021. 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. Springer; p. 154–162.

Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L. 2018. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE. p. 691–699.

Kranzfelder M, Schneider A, Fiolka A, Schwan E, Gillen S, Wilhelm D, Schirren R, Reiser S, Jensen B, Feussner H. 2013. Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology. journal of surgical research. 185(2):704–710.

Kurmann T, Marquez Neila P, Du X, Fua P, Stoyanov D, Wolf S, Sznitman R. 2017. Simultaneous recognition and pose estimation of instruments in minimally invasive surgery. In: International conference on medical image computing and computer-assisted intervention. Springer. p. 505–513.

Lalys F, Riffaud L, Bouget D, Jannin P. 2011. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Transactions on Biomedical Engineering. 59(4):966–976.

Liu YC, Ma CY, He Z, Kuo CW, Chen K, Zhang P, Wu B, Kira Z, Vajda P. 2021. Unbiased teacher for semi-supervised object detection. In: Proceedings of the International Conference on Learning Representations (ICLR).

Mishra K, Sathish R, Sheet D. 2017. Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In: Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. p. 58–65.
Raju A, Wang S, Huang J. 2016. M2cai surgical tool detection challenge report. In: Workshop and Challenges on Modeling and Monitoring of Computer Assisted Intervention (M2CAI), Athens, Greece, Technical report. p. 1–4.
Reiter A, Allen PK, Zhao T. 2012. Articulated surgical tool detection using virtually-rendered templates. In: Computer Assisted Radiology and Surgery (CARS). p. 1–8.
Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. 2015. Imagenet large scale visual recognition challenge. International journal of computer vision. 115(3):211–252.
Sahu M, Mukhopadhyay A, Szengel A, Zachow S. 2016. Tool and phase recognition using contextual cnn features. arXiv preprint arXiv:161008854.
Sarikaya D, Corso JJ, Guru KA. 2017. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE transactions on medical imaging. 36(7):1542–1549.
Shi P, Zhao Z, Hu S, Chang F. 2020a. Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network. IEEE Access. 8:228853–228862.
Shi P, Zhao Z, Hu S, Chang F. 2020b. Real-time surgical tool detection in minimally invasive surgery based on attention-guided convolutional neural network. IEEE Access. 8:228853–228862.
Sohn K, Berthelot D, Carlini N, Zhang Z, Zhang H, Raffel CA, Cubuk ED, Kurakin A, Li CL. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems. 33:596–608.
Speidel S, Benzko J, Krappe S, Sudra G, Azad P, Müller-Stich BP, Gutt C, Dillmann R. 2009. Automatic classification of minimally invasive instruments based on endoscopic image sequences. In: Medical Imaging 2009: Visualization, Image-Guided Procedures, and Modeling; vol. 7261. International Society for Optics and Photonics. p. 72610A.
Twinanda AP, Mutter D, Marescaux J, de Mathelin M, Padoy N. 2016a. Single-and multi-task architectures for tool presence detection challenge at m2cai 2016. arXiv preprint arXiv:161008851.
Twinanda AP, Shehata S, Mutter D, Marescaux J, De Mathelin M, Padoy N. 2016b. Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging. 36(1):86–97.
Van Engelen JE, Hoos HH. 2020. A survey on semi-supervised learning. Machine Learning. 109(2):373–440.
Vardazaryan A, Mutter D, Marescaux J, Padoy N. 2018. Weakly-supervised learning for tool localization in laparoscopic videos. In: Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis. Springer; p. 169–179.
Wang S, Xu Z, Yan C, Huang J. 2019. Graph convolutional nets for tool presence detection in surgical videos. In: International Conference on Information Processing in Medical Imaging. Springer. p. 467–478.
Ward TM, Mascagni P, Ban Y, Rosman G, Padoy N, Meireles O, Hashimoto DA. 2021. Computer vision in surgery. Surgery. 169(5):1253–1256.
Wu Y, Kirillov A, Massa F, Lo WY, Girshick R. 2019. Detectron2; [`https://github.com/facebookresearch/detectron2`].
Xu M, Zhang Z, Hu H, Wang J, Wang L, Wei F, Bai X, Liu Z. 2021. End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. p. 3060–3069.
Yang Y, Zhao Z, Shi P, Hu S. 2021. An efficient one-stage detector for real-time surgical tools detection in robot-assisted surgery. In: Annual Conference on Medical Image Understanding and Analysis. Springer. p. 18–29.
Yoon J, Lee J, Park S, Hyung WJ, Choi MK. 2020. Semi-supervised learning for instrument detection with a class imbalanced dataset. In: Interpretable and annotation-efficient learning

for medical image computing. Springer; p. 266–276.
Zhang B, Wang S, Dong L, Chen P. 2020. Surgical tools detection based on modulated anchoring network in laparoscopic videos. IEEE Access. 8:23748–23758.