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

Carlos Aparicio-Viveros[1], Mansoor Ali[1], Gilberto Ochoa-Ruiz[1], Sharib Ali[2]
Tecnologico de Monterrey[1], University of Leeds[2]
carlos.apaviv@outlook.com, {A01753093, gilberto.ochoa}@tec.mx, s.s.ali@leeds.ac.uk

Abstract

Surgical tool detection in minimally invasive surgery is an essential part of computer-assisted interventions. Current approaches are mostly based on supervised methods requiring large annotated datasets. However labelled datasets are often scarcely available. Semi-supervised learning (SSL) has recently emerged as a viable alternate and shown promise to produce models having competitive performance with supervised methods. Therefore, in this paper we introduce an 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 labelled data which initialises the Teacher-Student joint learning, where the Student is trained on Teacher-generated pseudo labels from unlabelled data. We also 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-background region. Our results on m2cai16-tool-locations dataset indicate the superiority of our approach on different supervised data settings (1%, 2%, 5% and 10% of annotated data) where our model achieves overall improvements of 8%, 12% and 27% in mean average precision on 1% labelled data over the state-of-the-art SSL methods and the supervised baseline, respectively.

1. Introduction

Recently, Computer-Assisted Intervention (CAI) systems capable of performing effectively the sub-tasks such as surgical phase recognition or surgical tool recognition and detection are getting increased attention, since the development of these task-based automated approaches can ensure improved surgical care (1).

In the realm of surgical tool detection, deep learning-based approaches have increasingly become a focal point of interest in recent years. However, the adoption of complex deep learning (DL) models is hampered by the need for extensive, precisely annotated datasets. Acquiring such datasets is a time-consuming task, affected by susceptibility to intra and inter-observer bias in annotations. Consequently, only a few labeled surgical tool datasets are publicly available (3; 6), hindering the development of robust and generalizable deep architectures for surgical instrument detection.

Despite the potential of deep learning in transforming surgical tool detection, existing methods predominantly rely on supervised learning approaches (3; 12), while only a few weakly supervised methods focusing on tool presence detection (10). This limitation underscores the need for innovative approaches that can circumvent the dependency on extensively annotated datasets.

To address these challenges, the annotation cost could be greatly mitigated by exploiting unlabeled data through efficient semi-supervised learning (SSL) frameworks. One way to do this is by leveraging unlabeled data for label prediction by first training a Teacher model on a pretext task to generate pseudo labels, then transferring this knowledge to a Student network for the primary task (4; 11). This process enriches the Student network's predictive accuracy with insights from both labeled and unlabeled data.

Although SSL has shown promising outcomes in improving model performance and is receiving growing attention of the computer vision research community (9), most of these advances are in the domain of image classification. Furthermore, SSL for object detection had been traditionally implemented by adapting state-of-the-art image classification methods such as (7) to object detection. However, the transition from image classification to object detection with Semi-Supervised Learning (SSL) methods is challenging due to critical differences between these tasks. Object detection is particularly hampered by imbalances between foreground-background and within classes, which can compromise pseudo-labeling efficacy and introduce biases towards dominant classes. Thus, applying SSL techniques designed for classification directly to detection risks exacerbating class imbalances and may result in significant overfitting.

To overcome these issues, we propose a jointly trained Teacher-Student model on m2cai16-tool-locations dataset (3) 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 and, hence, overall performance improvement. Additionally, we propose a multi-class distance and margin-based classification loss in the region-of-interest (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. Our proposed pipeline outperforms supervised baseline and other SOTA semi-supervised methods in terms of classification and localisation performance.

2. Data

2.1. Dataset

In this work, we use an extended version of the m2cai16-tool dataset, originally released for the M2CAI 2016 Tool Presence Detection Challenge (8), now including the m2cai16-tool-locations dataset (3). This extension provides spatial bounding box annotations for 7 classes: grasper, bipolar, hook, scissors, clipper, irrigator, and specimen bag, across a total of 2812 annotated frames. Annotations were performed under the supervision and spot-checking of clinical experts, with dataset splits of 80%, 10%, and 10% for training, validation, and testing, respectively.

2.2. Data Augmentation

We have used two data augmentation strategies in this work, which we refer as weak and strong augmentations. 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 and cutout patches (2). For the complete description of data augmentation with parameter values, please refer to Liu et al. (4).

3. 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. 1. The first stage is the initialization stage (section 3.1), while the second is the Teacher-Student joint learning mechanism (section 3.2).

3.1. 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 (θ) 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)).

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

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.

3.2. 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.

3.3. Student learning and Teacher update scheme

We tackle the pseudo-label noise problem which may cause severe performance degradation (7) by confidence thresholding (τ). We address the duplicated box predictions problem by applying class-wise non-maximum suppression (NMS) before a confidence thresholding step. As

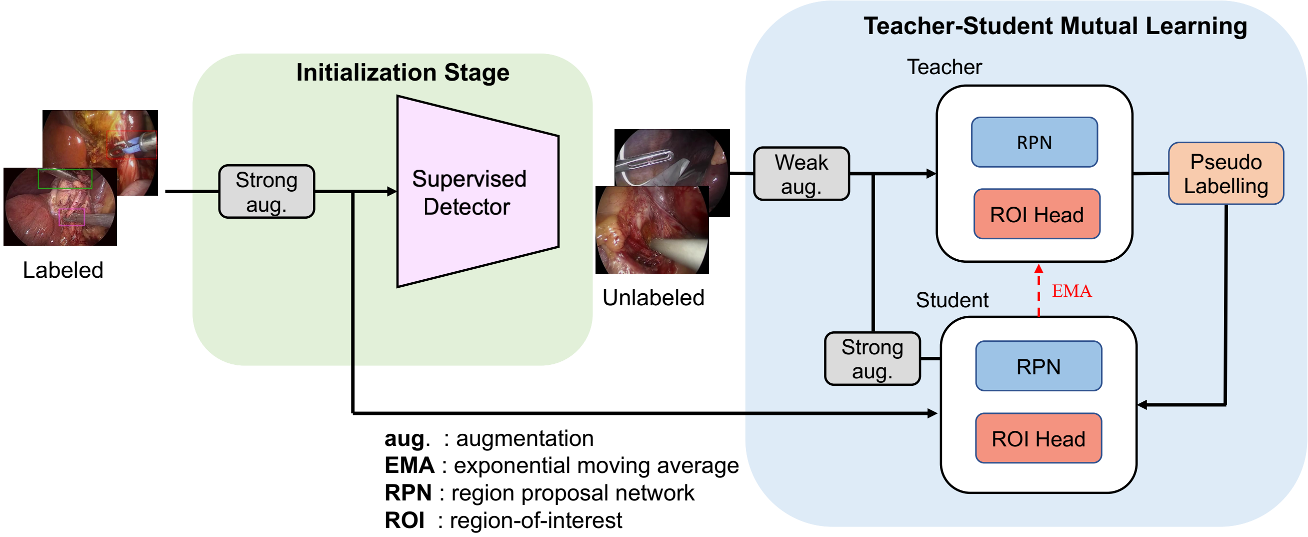

Figure 1. Overview of the proposed Surgical tool detection model. It consists of two modules: (1) An initialisation module, where a supervised model makes use of strongly augmented labelled data, and (2) A Teacher-Student mutual learning module, where the Student is trained with strongly augmented unlabelled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

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 θ_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 γ is the learning rate and λ_u is the 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 λ_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 α is the EMA rate, and $\theta_{\mathcal{T}}$, $\theta_{\mathcal{S}}$ are the network weights for Teacher and Student.

3.4. Logistic loss with added margin and distance penalisation

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 (5). We target the foreground-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 ρ and background β 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 σ 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.

4. Experiments and results

4.1. 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 (τ) of 0.7, regularization co-efficient for unsupervised loss (λ_u) of 0.2 and EMA rate (α) of 0.9996. We use *WarmupMultiStepLR* as a learning rate (α) 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 use a batch size of 8 (4 labeled images and 4 unlabeled images) throughout the experiments. We performed network training using 4 graphical processing units (GPUs) on NVIDIA Tesla P100-SXM2-16GB system.

4.2. 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 (4) with both CrossEntroppy and focal losses. and SoftTeacher (11). We report results in terms of mAP 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).

Our experiments on m2cai16-tool dataset show the effectiveness of our model in terms of mAP on various supervision protocols against the supervised setting and the SOTA semi-supervised models.

4.3. Discussion and Conclusion

We demonstrate that our proposed approach performs favourably against the SOTA semi-supervised models proposed by (4) and (11) . 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 (11) 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 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 introduction of loss function which effectively increases the foreground-background distance and helps in improving detection performance.

	1% Labelled data		
Method	mAP_{50}	$mAP_{50:95}$	mAP_{75}
Supervised	23.578	7.673	2.322
Unbiased Teacher*	34.374	14.145	7.855
Unbiased Teacher**	42.382	18.008	11.387
SoftTeacher (11)	38.421	13.556	6.623
Ours	50.632	20.094	12.713
	2% Labelled data		
Method	mAP_{50}	$mAP_{50:95}$	mAP_{75}
Supervised	47.140	18.609	9.480
Unbiased Teacher*	71.608	31.752	20.479
Unbiased Teacher**	72.416	31.490	21.446
SoftTeacher (11)	60.366	25.421	14.767
Ours	72.341	32.311	21.614
	5% Labelled data		
Method	mAP_{50}	$mAP_{50:95}$	mAP_{75}
Supervised	71.082	32.249	21.995
Unbiased Teacher*	84.721	42.269	32.826
Unbiased Teacher**	82.592	40.393	30.735
SoftTeacher (11)	83.211	38.857	26.643
Ours	84.427	42.392	33.376
	10% Labelled data		
Method	mAP_{50}	$mAP_{50:95}$	mAP_{75}
Supervised	80.193	38.640	30.625
Unbiased Teacher*	92.981	47.369	41.049
Unbiased Teacher**	90.353	45.972	45.103
SoftTeacher (11)	89.362	42.717	41.522
Ours	90.250	46.886	46.234

[*] Unbiased Teacher with focal loss (4)
[**] Unbiased Teacher with cross entropy loss (4)

Table 1. Experimental results with ResNet50-FPN as backbone

Furthermore, the proposed framework performs much better on mAP_{75} in all settings consistently which indicates improved localisation performance. If we compare the performance of our model on *mAP* at 50:95 and 75 IoU thresholds on the 2%, 5% and 10% labeled data settings, we observe that our model consistently gives superior performance. This validates the effectiveness of our method on both classification and localization performance.

In this work, 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.

In this paper, we addressed a lack of annotated data problem in surgical domain for the first time by proposing 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.

References

[1] David Bouget, Max Allan, Danail Stoyanov, and Pierre Jannin. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Medical image analysis*, 35:633–654, 2017. 1

[2] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[3] Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. 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)*, pages 691–699. IEEE, 2018. 1, 2

[4] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 4

[5] Kaustuv Mishra, Rachana Sathish, and Debdoot Sheet. 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*, pages 58–65, 2017. 3

[6] Duygu Sarikaya, Jason J Corso, and Khurshid A Guru. 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, 2017. 1

[7] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33:596–608, 2020. 1, 2

[8] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016. 2

[9] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020. 1

[10] Armine Vardazaryan, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 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*, pages 169–179. Springer, 2018. 1

[11] 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. 1, 4

[12] Beibei Zhang, Shengsheng Wang, Liyan Dong, and Peng Chen. Surgical tools detection based on modulated anchoring network in laparoscopic videos. *IEEE Access*, 8:23748–23758, 2020. 1