scientific reports

Check for updates

OPEN

SegMatch: semi-supervised surgical instrument segmentation

Meng Wei[1]✉, Charlie Budd[1], Luis C. Garcia-Peraza-Herrera[1], Reuben Dorent[3], Miaojing Shi[2]✉ & Tom Vercauteren[1]

Surgical instrument segmentation is recognised as a key enabler in providing advanced surgical assistance and improving computer-assisted interventions. In this work, we propose SegMatch, a semi-supervised learning method to reduce the need for expensive annotation for laparoscopic and robotic surgical images. SegMatch builds on FixMatch, a widespread semi-supervised *classification* pipeline combining consistency regularization and pseudo-labelling, and adapts it for the purpose of *segmentation*. In our proposed SegMatch, the unlabelled images are first weakly augmented and fed to the segmentation model to generate pseudo-labels. In parallel, images are fed to a strong augmentation branch and consistency between the branches is used as an unsupervised loss. To increase the relevance of our strong augmentations, we depart from using only handcrafted augmentations and introduce a trainable adversarial augmentation strategy. Our FixMatch adaptation for segmentation tasks further includes carefully considering the equivariance and invariance properties of the augmentation functions we rely on. For binary segmentation tasks, our algorithm was evaluated on the MICCAI Instrument Segmentation Challenge datasets, Robust-MIS 2019 and EndoVis 2017. For multi-class segmentation tasks, we relied on the recent CholecInstanceSeg dataset. Our results show that SegMatch outperforms fully-supervised approaches by incorporating unlabelled data, and surpasses a range of state-of-the-art semi-supervised models across different labelled to unlabelled data ratios.

Automatic visual understanding in laparoscopic and robotic surgical videos is crucial for enabling autonomous surgery and providing advanced surgical support to clinical teams. Within this field, instrument segmentation, as shown in Fig. 1, is a fundamental building block. Example use cases include automatic surgical skill assessment, placing informative overlays on the display, performing augmented reality without occluding instruments, intra-operative guidance systems, surgical workflow analysis, visual servoing, and surgical task automation[1]. Surgical instrument segmentation has advanced rapidly from traditional methods[2,3], to modern deep learning-based methods[4–7]. Given the significance of this task, open datasets have been released, in particular, through the organisation of computational challenges such as Robust-MIS 2019[1].

Most studies in this area exploit fully-supervised learning, the performance of which scales with the amount of labelled training data. However, given the required expertise to provide accurate manual segmentation, the collection of annotations is costly and time-consuming. It is thus unsurprising that in comparison to industry standards for natural images, no large-scale annotated datasets for surgical tool segmentation currently exist. This leads to significant barriers in establishing the robustness and precision needed to deploy surgical instrument segmentation in the clinic. To tackle this challenge, a number of weak supervision approaches[8–12] have been proposed which take advantage of unlabelled data or image-level labels as these are easier to capture. While interesting conceptually, weak supervision for surgical instrument segmentation has not yet been demonstrated in practice to generalize to the largest open-access surgical datasets and achieve the required accuracy for deployment in the operating theatre.

In this work, we propose a new semi-supervised surgical instrument segmentation framework, termed as SegMatch, building upon the widely adopted semi-supervised image *classification* pipeline, FixMatch[13]. During training, FixMatch processes unlabelled input images through two concurrent paths which implement weak (*e.g.* image flip and rotation) and strong (*e.g.* significant photometric changes) augmentations respectively. Augmented images from both paths are then fed to a shared backbone prediction network. For regions of high confidence, the prediction from the weakly augmented image serves as a pseudo-ground-truth against which the strongly augmented image is compared.

[1]School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK. [2]College of Electronic and Information Engineering, Tongji University, Shanghai, China. [3]Harvard Medical School, Harvard University, Boston, MA, USA. ✉email: meng.wei@kcl.ac.uk; mshi@tongji.edu.cn

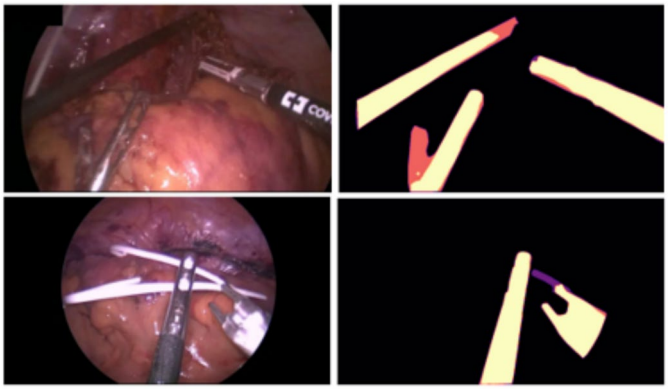


Fig. 1. Representative sample images from Robust-MIS 2019 of laparoscopic surgery (left) and state-of-the-art instrument segmentation results (right). True positive (yellow), true negative (black), false positive (purple), and false negative (red).

In order to adapt FixMatch to the segmentation task, we make a seemingly small but critical first contribution by changing the augmentation paths in SegMatch. For classification tasks, networks are expected to be invariant with respect to all types of augmentation within a certain range tailored to the specific requirements of the task. In contrast, for segmentation tasks, networks are expected to be invariant with respect to photometric transformations (*e.g.* contrast, brightness, hue changes) but equivariant with respect to spatial transformations (*e.g.* rotations, translations, flips). In SegMatch, spatial transformations that are used as augmentations are inverted after the network prediction.

Our second and main contribution to SegMatch lies in the inclusion of a learned strong augmentation strategy. Promoting prediction consistency between the weakly augmented and strongly augmented branches is what helps FixMatch and SegMatch learn from unlabelled data and generalise better. Yet, there is no guarantee that the fixed, hand-crafted set of strong augmentation types suggested in previous work[13] is optimal. In fact, once the network learns to be sufficiently invariant/equivariant with respect to the fixed set of augmentation, the information learned from the unlabelled data would have saturated. In order to guide the model to learn continuously as the training progresses, we introduce an adversarial augmentation scheme to generate strongly augmented data. We rely on the established iterative fast gradient sign method (I-FGSM)[14] and integrate it into our training loop to dynamically adapt the strong augmentation as we learn from the data.

We conduct extensive experiments on the Robust-MIS 2019[1] and EndoVis 2017[15] datasets for binary segmentation tasks, as well as on the CholecInstanceSeg[16] dataset for multi-class segmentation task. Our study demonstrated that using a ratio of 1:9 or 3:7 of labelled data to unlabelled data within a dataset allowed our model to achieve a considerably higher mean Dice score compared to state-of-the-art semi-supervised methods. Our method shows a significant improvement in surgical instrument segmentation compared to existing fully-supervised methods by utilizing a set of 17K unlabelled images available in the Robust-MIS 2019 dataset in addition to the annotated subset which most competing methods have exploited.

The key contributions of this work are:

- We adapt the representative semi-supervised classification pipeline with segmentation-specific transformations for weak and strong augmentations and consistency regularization.
- We propose a trainable adversarial augmentation strategy to improve the relevance of strong augmentations, ensuring the model learns robustly by leveraging both handcrafted and dynamic augmentations.
- Extensive experiments on the EndoVis 2017 and CholecInstanceSeg datasets demonstrate superior performance by our SegMatch compared to existing semi-supervised models for both binary and multi-class segmentation tasks.

Related work

Semi-supervised learning

Pseudo-labelling[17,18] is a representative approach for semi-supervised learning. A model trained on labelled data is utilized to predict pseudo-labels for unlabelled data. This in turn provides an extended dataset of labelled and pseudo-labelled data for training. Consistency regularization[19,20] is also a widespread technique in semi-supervised learning. There, an auxiliary objective function is used during training to promote consistency between several model predictions, where model variability arises from techniques such as weight smoothing or noise injection.

Berthelot et al.[21] introduced MixMatch to incorporate both consistency regularization and the Entropy Minimization strategy of Grandvalet et al.[22] into a unified loss function for semi-supervised image classification. Aiming at providing a simple yet strong baseline, Sohn et al.[13] introduced FixMatch to combine consistency regularization and pseudo-labelling and achieve state-of-the-art performance on various semi-supervised learning benchmarks.

Adapting semi-supervised learning for image segmentation tasks requires dedicated strategies that account for the fact that labels are provided at the pixel level. Previous works explored the adaptation to semantic segmentation of classical semi-supervised learning including pseudo-labelling[23,24], and consistency regularization[25,26]. Ouali et al.[27] proposed cross-consistency training (CCT) to force the consistency between segmentation predictions of unlabelled data obtained from a main decoder and those from auxiliary decoders. Similarly, Chen et al.[28] exploited a novel consistency regularization approach called cross pseudo-supervision such that segmentation results of variously initialised models with the same input image are required to have high similarity. Previous work also investigated the use of generative models to broaden the set of unlabelled data from which the segmentation model can learn[29]. More recently, Wang et al.[30] proposed a contrastive learning approach for semi-supervised dense prediction, introducing adversarial negatives and an auxiliary classifier designed to track and exclude potential false negatives.

Despite such advances, current semi-supervised semantic segmentation models derived from classification models have not yet demonstrated the performance gains observed with FixMatch for classification. We hypothesise that an underpinning reason is that they do not adequately address the issue of transformation equivariance and invariance and do not exploit modern augmentation strategies as efficiently as FixMatch.

Surgical instrument segmentation

The majority of surgical instrument segmentation works are supervised methods[6,31–34]. Numerous studies have explored different methods to improve the accuracy of surgical instrument segmentation. For instance, Islam et al.[6] took advantage of task-aware saliency maps and the scan path of instruments in their multitask learning model for robotic instrument segmentation. By using optical flow, Jin et al.[33] derived the temporal prior and incorporated it into an attention pyramid network to improve the segmentation accuracy. Gonzalez et al.[31] proposed an instance-based surgical instrument segmentation network (ISINet) with a temporal consistency module that takes into account the instance's predictions across the frames in a sequence. Seenivasan et al.[34] exploited global relational reasoning for multi-task surgical scene understanding which enables instrument segmentation and tool-tissue interaction detection. TraSeTR[35] utilizes a Track-to-Segment Transformer that intelligently exploits tracking cues to enhance surgical instrument segmentation. Recently, large models have started to be used. Wei et al.[36] for example introduced an adapter network that integrates pre-trained knowledge from foundation models into a lightweight convolutional model, enhancing both robustness and performance in surgical instrument segmentation. Yet, the use of unlabelled task-specific data for surgical tool segmentation remains relatively untapped.

A relatively small number of works exploited surgical instrument segmentation with limited supervision. Jin et al.[33] transferred predictions of unlabelled frames to that of their adjacent frames in a temporal prior propagation-based model. Liu et al.[10] proposed an unsupervised approach that relies on handcrafted cues including colour, object-ness, and location to generate pseudo-labels for background tissues and surgical instruments respectively. Liu et al.[37] introduced a graph-based unsupervised method for adapting a surgical instrument segmentation model to a new domain with only unlabelled data. We note that most of the existing works with limited supervision focus on exploring domain adaptation or generating different types of pseudo-labels for surgical tool segmentation models and do not exploit a FixMatch-style semi-supervised learning.

Adversarial learning for improved generalisation

Deep neural networks (DNNs) have been found to be vulnerable to some well-designed input samples, known as adversarial samples. Adversarial perturbations are hard to perceive for human observers, but they can easily fool DNNs into making wrong predictions. The study of adversarial learning concerns two aspects: (1) how to generate effective adversarial examples to attack the model[38]; (2) how to develop efficient defence techniques to protect the model against adversarial examples[39]. A model which is robust to adversarial attacks is also more likely to generalise better[40]. As such, we hypothesise that adversarial methods may be of relevance for semi-supervised learning.

For adversarial attacks, the earliest methods[38,41] rely on the gradient of the loss with respect to the input image to generate adversarial perturbations. For instance, fast gradient sign attack (FGSM)[41] perturbs the input along the gradient direction of the loss function to generate adversarial examples. Tramer et al.[42] improved the FGSM by adding a randomization step to escape the non-smooth vicinity of the data point for better attacking. The basic iterative method (BIM)[14], which is also referred to as iterative FGSM (I-FGSM), improves FGSM by performing multiple smaller steps to increase the attack success rate. Carlin et al.[43] introduced a method now referred to as C&W which solves a constrained optimisation problem minimising the size of the perturbation while ensuring a wrong prediction after perturbation. Recently, Rony et al.[44] proposes a more efficient approach to generate gradient-based attacks with low L2 norm by decoupling the direction and norm (DDN) of the adversarial perturbation. For adversarial defence, adversarial training[45] is a seminal method which generates adversarial examples on the fly and trains the model against them to improve the model's robustness. Other defence methods include relying on generative models[46] and leveraging the induced randomness to mitigate the effects of adversarial perturbations in the input domain[47].

Although the adversarial attack has the potential to enhance the performance and robustness of deep learning models, it has not been yet applied to semi-supervised learning methods for semantic segmentation. As we show later in this work, adversarial attacks can indeed be used to effectively augment unlabelled images used for consistency optimization.

Method

Proposed SegMatch algorithm: overview

Our proposed SegMatch algorithm adapts the state-of-the-art semi-supervised image *classification* framework, FixMatch[13], to semi-supervised semantic *segmentation*. Our application primarily targets the segmentation of surgical instruments but also has the potential to be utilized for various other semantic segmentation tasks. We follow the basic architecture of FixMatch as illustrated in Fig. 2. During training, SegMatch uses a supervised pathway and an unsupervised pathway which share the same model parameters θ for segmentation prediction. For the supervised pathway, the prediction from the labelled image is classically optimized against its ground truth using a standard supervised segmentation loss. For the unsupervised pathway, given an unlabelled image, we follow FixMatch and process the input image through two concurrent paths, which implement weak and strong augmentations respectively. The augmented images are then fed to the model to obtain two segmentation proposals. The output from the weak augmentation branch is designed to act as the pseudo-ground-truth for that from the strong augmentation branch.

As detailed in Sect. Weak augmentations, equivariance, pseudo-labels, segmentation models are typically expected to be invariant with respect to bounded photometric transformations and equivariant with respect to spatial transformations of the input images. We use simple spatial transformations for our weak augmentations and propose to apply the inverse spatial transformation after the network prediction to handle spatial transformation equivariance. We use photometric transformations for our strong augmentations and exploit a pixel-wise loss promoting consistency between the segmentation outputs of the two branches.

Given the complexity of determining the suitable range of parameters for hand-crafted strong augmentation, we propose a solution that involves a learning-based approach. As detailed in Sect. Trainable strong augmentations, in order to gradually increase the difficulty of the prediction from the strongly augmented branch, we introduce a learnable adversarial augmentation scheme in the strong augmentation branch.

Weak augmentations, equivariance, pseudo-labels

Equivariance and invariance in SegMatch

We start by introducing the notion of equivariance and invariance illustrated in Fig. 3. Let us consider a function f_θ (standing for the neural network, θ being the parameters of the network) with an input x (standing for the input image), and a class of transformation functions $\mathcal{G}$ (standing for a class of augmentations). If applying $g \in \mathcal{G}$ (standing for a specific augmentation instance) to the input x also reflects on the output of f_θ, that is if $f_\theta(g(x)) = g(f_\theta(x))$, then the function f_θ is said to be equivariant with respect to transformations in $\mathcal{G}$ (see

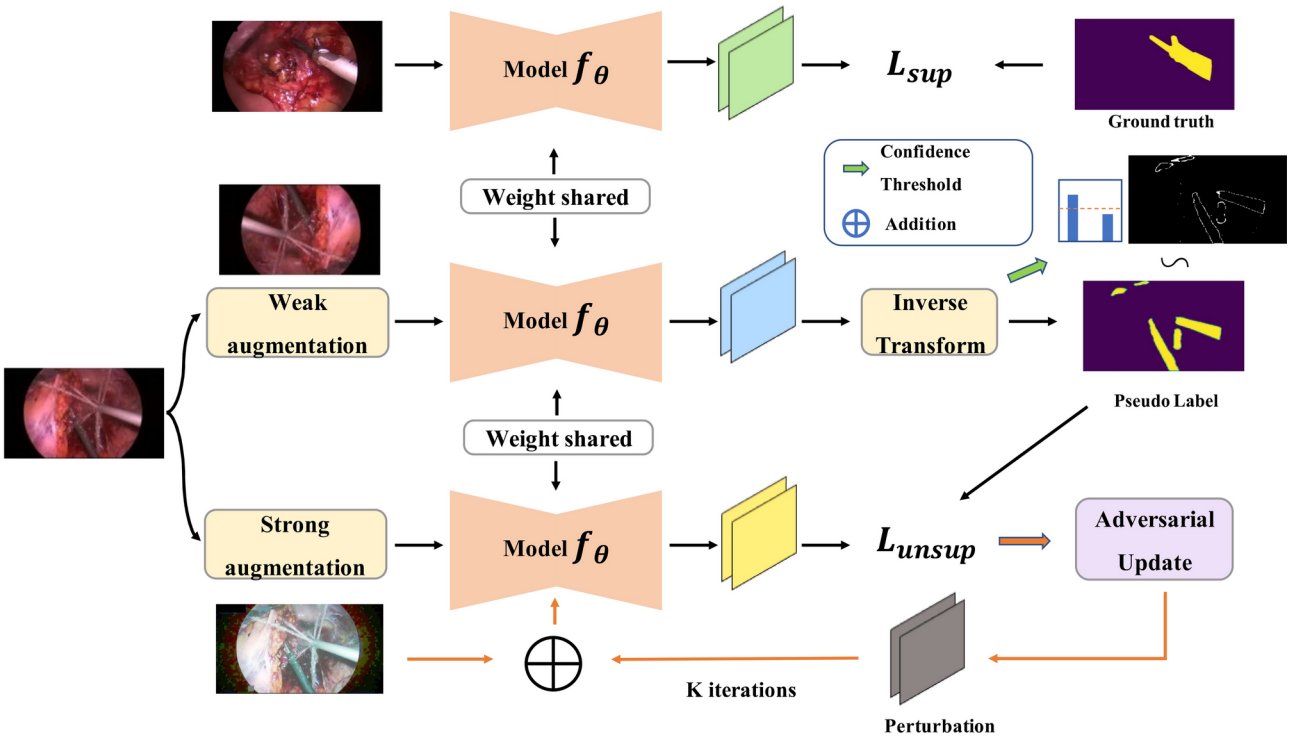


Fig. 2. SegMatch training process structure. The top row is the fully-supervised pathway which follows the traditional segmentation model training process. The two bottom rows form the unsupervised learning pathway, where one branch uses a weakly augmented image fed into the model to compute predictions, and the second branch obtains the model prediction via strong augmentation for the same image. The model parameters are shared across the two pathways. The hand-crafted photometric augmentation methods are used to initialize the strong augmented image, which is further perturbed by an adversarial attack (I-FGSM) for K iterations.

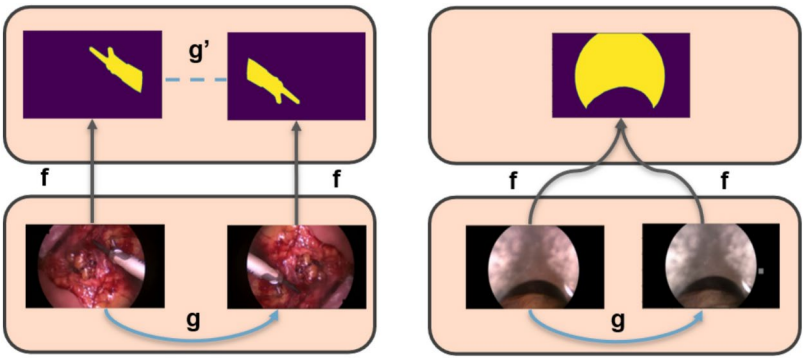


Fig. 3. Equivariance (left) and invariance (right) properties for an image augmented under different types of augmentations: spatial (left) or photometric (right).

Fig. 3-Left). Conversely, if applying $g \in \mathcal{G}$ to the input x does not affect the output of f_θ, that is $f_\theta(g(x)) = f_\theta(x)$, then the function f_θ is said to be invariant with respect to transformations in $\mathcal{G}$ (see Fig. 3-Right).

For the classification task, the model is expected to be invariant to all the augmentation strategies. In contrast, given a segmentation model, spatial transformations on the input image should reflect on the output segmentation map while photometric transformations should not. Segmentation models should thus be equivariant with respect to spatial transformations and invariant for photometric transformation. In FixMatch, weak augmentations only comprise simple spatial transformations such as rotation, and flip, which preserve the underlying structure of the image. Meanwhile, strong augmentations only comprise photometric transformations such as posterizing and sharpness changes as provided by the RandAugment[48] algorithm, which shifts the intensity distribution of the original image.

Inverting transformations from the weak augmentations

Similar to FixMatch, given an input unlabelled image, we randomly select one simple spatial transformation, *i.e.* rotation, flip, crop or resize, to apply to it in the weak augmentation branch. In order for our SegMatch to take advantage of a consistency loss between the weak augmentation branch where spatial transformations are used (with expected equivariance of the segmentation) and the strong augmentation branch where photometric transformations are used (with expected invariance of the segmentation), we deploy an inverse spatial transformation on the output of the segmentation model in the weak augmentation branch.

Given an unlabelled image x^u, we denote its weak augmentation as $\omega_e(x^u)$. It is fed to the network f_θ to obtain the segmentation output $f_\theta(\omega_e(x^u))$. We apply the inverse transformation ω_e^{-1} to $f_\theta(\omega_e(x^u))$, *i.e.* $\omega_e^{-1}(f_\theta(\omega_e(x^u)))$. This addresses the equivariance expectation and allows for the output of the weak augmentation branch to be easily compared with the segmentation output from the strongly augmented image.

Soft pseudo-label generation

Following the inverse transformation, we obtain a segmentation prediction in logit space $p_w = \omega_e^{-1}(f_\theta(\omega_e(x^u)))$. For the i-th pixel in p_w, *i.e.* p_{w_i}, a pseudo-label is extracted by applying a sharpened softmax:

$$\widetilde{y_i} = \text{Sharpen}(\text{Softmax}(p_{w_i}), T) \tag{1}$$

where the distribution sharpening operation of[21] is defined as

$$\text{Sharpen}(d, T)_i := \frac{d_i^{\frac{1}{T}}}{\sum_{j=1}^{L} d_j^{\frac{1}{T}}} \tag{2}$$

with T being a temperature hyper-parameter. The sharpening operation allows us to control the level of confidence in the resulting probabilities.

Trainable strong augmentations

To tackle the generalization problem typically faced by convolutional neural networks, previous work has employed strong augmentations techniques[13,21,49]. However, these augmentations are hand-crafted and designing realistic strong augmentations to bridge large and complex domain gaps is challenging[50]. This challenge is further exacerbated in segmentation tasks which are highly sensitive to pixel-level perturbations due to their pixel-level prediction nature[51]. For these reasons, despite utilizing powerful hand-crafted augmentations during training, existing methods still demonstrate limited generalization capabilities. In this section, we propose to tackle these key limitations by learning to perform data augmentation using adversarial perturbation during model training.

Strong augmentation initialization
Rather than completely replacing the strong augmentation approach in FixMatch, we build on it to serve as initialization which will be further perturbed. We chose the strong augmentations from the operations in RandAugment[48] based on two criteria. First, we focus on photometric transformations only as these satisfy the invariance expectation and do not require the use of an inverse transformation as discussed in Sect. Weak augmentations, equivariance, pseudo-labels. Second, we select rather basic transformations that provide visually plausible augmentations, thereby leaving the more complex changes to the trainable refinement. More specifically, our initial augmentation for the strong augmentation branch is a composition of three transformations randomly chosen from a collection of handcrafted photometric augmentation strategies. These include adjustments to contrast, brightness, colour, and sharpness, as well as the addition of random noise, posterization, and solarization. The strength of the individual transformations is chosen according to predefined hyper-parameters.

Adversarial augmentation approach
As a simple yet powerful adversarial method, we decide to use the iterative fast gradient sign method (I-FGSM)[14], which applies multiple gradient updates by iterative small steps.

I-FGSM is based on FGSM which provides an adversarial perturbation to the input image in a single gradient-based operation. In FGSM, the direction of the perturbation is computed from the gradient of the loss with respect to the input image. The magnitude of the gradient is discarded by keeping only the sign along each dimension. A scaling factor is applied to keep the perturbation small. To compute a more refined perturbation, I-FGSM applies FGSM multiple times in smaller steps. The perturbations are clipped to make sure they are in the ϵ-neighbourhood to the original image. The I-FGSM equation for iteration $k+1$ out of K is as follows:

$$x^s_{k+1} = \mathrm{Clip}_{x^s_0,\epsilon}\{x^s_k + \frac{\epsilon}{K} \cdot \mathrm{Sign}(\nabla_{x^s_k}(L_u(f_\theta(x^s_k), \widetilde{y}))\} \tag{3}$$

where x^s_0 represents the initial strongly-augmented image; $\widetilde{y}$ is the pseudo-label obtained from the weak augmentation branch; $\mathrm{Clip}\{\cdot\}$ is the clipping function which applies to every pixel in the perturbation image to limit the difference between x^s_K and x^s_0 and keep it within an ϵ-neighbourhood; and L_u is the unsupervised loss function defined in Eq. (5). The magnitude of the perturbation ϵ and the number of I-FGSM steps K are hyperparameters to adjust the quality of the adversarial approach.

Loss functions in SegMatch

In this work, the training objective for the supervised pathway is the standard pixel-wise cross-entropy loss (l_{CE}) combined with Dice loss ($l_{DSC} = 1 - DSC$), where *DSC* represents the soft Dice coefficient:

$$L_s = \frac{1}{|\mathcal{D}^l|} \sum_{x^l \in \mathcal{D}^l} \left(l_{DSC}\left(y^l, f_\theta(x^l)\right) + \frac{1}{N} \sum_{i=0}^{N-1} l_{CE}\left(y^l_i, f_\theta(x^l)_i\right) \right) \tag{4}$$

where x^l is a labelled input from the labelled set $\mathcal{D}^l$; y^l is the corresponding ground-truth label; x^l_i and y^l_i denote the ith pixel of x^l and y^l, respectively; and N is the number of pixels in x^l.

The training objective for the unsupervised pathway is a cross-entropy loss calculated between a subset of confident pixel-level pseudo-labels $\widetilde{y_i}$ stemming from the weak augmentation branch and the output probability p_i from the strongly augmented image:

$$L_u = \frac{1}{|D_u|} \sum_{x^u \in D_u} \frac{1}{|N_v^{x^u}|} \sum_{i \in N_v^{x^u}} l_{CE}(\widetilde{y_i}, p_i) \tag{5}$$

where where x^u is an unlabelled input from the unlabelled set $\mathcal{D}^u$; c denotes a specific class; and $N_v^{x^u}$ is the set of pixel indices with confidence score of the most confident class $\max_c(p^c_{w_i})$ higher than or equal to a hyper-parameter threshold t, i.e. $N_v^{x^u} = \{i \mid \max_c(p^c_{w_i}) \geq t\}$.

The final loss is given by:

$$L = L_s + w(t)L_u \tag{6}$$

where, following Laine et al.[52], $w(t)$ is an epoch-dependent weighting function which starts from zero and ramps up along a Gaussian curve so that the supervised loss contributes to the total loss more at the beginning and the unsupervised loss increases contribution in subsequent training epochs.

Data flow

To provide an overall workflow for both the weak and strong augmentation branches in SegMatch, we present detailed pseudo-code illustrations for each component in Algorithms 1 and 2. Algorithm 1 outlines the workflow for the weak augmentation branch, while Algorithm 2 details the strong augmentation branch, which integrates hand-crafted strong augmentation initialization and adversarial augmentation.

Input: Unlabelled image X^u

Randomly select a weak augmentation from a predefined set of spatial transformations: ω_e

Apply weak augmentation to X^u: $\omega_e(x^u)$

Pass the augmented image through the segmentation network: $f_\theta(\omega_e(x^u))$

Apply the inverse spatial transformation to the output: $p_w = \omega_e^{-1}(f_\theta(\omega_e(x^u)))$

for each pixel p_{wi} in p_w **do**

Generate pseudo-label $\widetilde{y}_i$ by applying a sharpened softmax:

$$\widetilde{y}_i = \text{Sharpen}(\text{Softmax}(p_{wi}), T)$$

end for

Output: Model prediction of weak augmentation branch p_w

Pseudo label $\widetilde{y}$

Algorithm 1. Workflow for weak augmentation branch

Input: Unlabelled image X^u**; Pseudo label** $\widetilde{y}$**; Confidence threshold** t

Model prediction of weak augmentation branch p_w

Perturbation magnitude ε**; Number of steps for I-FGSM** K

Randomly select 3 photometric augmentation functions $\{a_1, a_2, a_3\}$ from RandAugment (set $\mathscr{A}$) and compose the transformations: $\mathscr{T} \leftarrow a_1 \circ a_2 \circ a_3$

Apply the composed transformation to the unlabelled image: $X_{stg} = \mathscr{T}(x^u)$

Initialization: $X_0^s = X_{stg}, i = 0$

repeat

Pass the augmented image through the segmentation network: $p = f_\theta(X_i^s)$

for each pixel $j \in p_w$ **do**

Identify the most confident class: $\max_c(p_{w_j}^c)$

Add pixel j to the confident pixel set $N_v^{x^u}$ if $max_c(p_{w_j}^c) \geq t$

end for

Compute the mean loss over the confident pixels $j \in N_v^{x^u}$:

$$L_u^{x^u} = \frac{1}{|N_v^{x^u}|} \sum_{j \in N_v^{x^u}} l_{CE}(\widetilde{y}_j, p_j)$$

Let $C_{\max}$ the maximum allowable pixel value, e.g. 255 if unnormalised images are used

Apply one step of FGSM:

$X_{i+1}^s \leftarrow X_i^s + \frac{\varepsilon}{K} \cdot \text{sign}(\bigtriangledown x(L_u^{x^u}))$

$X_{max} \leftarrow \max\{0, X_0^s - \varepsilon, X_{i+1}^s\}$

$X_{i+1}^s \leftarrow \min\{C_{max}, X_0^s + \varepsilon, X_{max}\}$

$i \leftarrow i + 1$

until $i = K - 1$

Output: Strongly augmented image X_K^s

Algorithm 2. Workflow for strong augmentation branch

Experimental setup

Dataset

Robust-MIS 2019

Robust-MIS 2019 is a laparoscopic instruments dataset including procedures in rectal resection, proctocolectomy, and sigmoid resection to detect, segment, and track medical instruments based on endoscopic video images[1]. The training data encompasses 10-second video snippets in the form of 250 consecutive endoscopic image frames and the reference annotation for only the last frame is provided. In total, 10,040 annotated images are available from a total of 30 surgical procedures from three different types of surgery.

As per the original challenge, the samples used for training were exclusively taken from the proctocolectomy and rectal resection procedures. These samples comprise a total of 5983 clips, with each clip having only one annotated frame while the remaining 249 frames are unannotated.

The testing set is divided into three phases as per the original challenge, where there was no patient overlap between the training and test datasets. *Stage 1*: 325 images from the proctocolectomy procedure with another 338 images from the rectal resection procedure. *Stage 2*: 225 images from the proctocolectomy procedure and 289 others from the rectal resection procedure. *Stage 3*: 2880 annotated images from sigmoid resection, an unknown surgery which only appears in the testing stage but not in the training stage.

EndoVis 2017
EndoVis 2017 is a robotic instrument image dataset captured from robotic-assisted minimally invasive surgery[15], which comprises a collection of 10 recorded sequences capturing abdominal porcine procedures. For the training phase, the first 225 frames of 8 sequences were captured at a rate of 2Hz and manually annotated with information on tool parts and types. The testing set consists of the last 75 frames from the 8 sequences used in the training data videos, along with 2 full-length sequences of 300 frames each, which have no overlap with the training phase. To prevent overlap between the training and test sets from the same surgical sequence, we followed the same split as described in ISINet[31]. This involved exclusively assigning the 2 full-length sequences for testing while keeping the training set intact with 225 frames from the remaining 8 sequences. Note that there are no additional unannotated images for EndoVis 2017.

CholecInstanceSeg
The CholecInstanceSeg dataset[16] is a comprehensive, instance-labeled collection featuring 41,933 frames from 85 unique video sequences. This dataset includes seven distinct tool categories: Grasper, Bipolar, Hook, Clipper, Scissors, Irrigator, and Snare. Each frame has been meticulously annotated, extracted from 85 laparoscopic cholecystectomy procedures in the Cholec80[53] and CholecT50[54] datasets. The dataset is divided into a training set with 55 sequences (26,830 frames), a testing set with 23 sequences (11,299 frames), and a validation set with 17 sequences (3,804 frames). In this work, we utilize the dataset solely for multi-class semantic segmentation and thus discard the instance-specific information.

Dataset usage for semi-supervised learning evaluation
Since the above challenges were designed for fully-supervised benchmarking, we make some adaptions to evaluate SegMatch with respect to competitive semi-supervised (and fully-supervised) methods.

The Robust-MIS 2019 dataset was split into training and testing sets based on the original challenge splits. The three original challenge testing stages were merged to form a single combined testing set, comprising 4057 images. For training, we started with the full set of 5983 labelled original challenge training images and the corresponding 17,617 unlabelled images. In our first experiments, we use only the 5983 labelled original challenge training images and keep only 10% or 30% of the annotations. This allows for a comparison with supervised methods having access to all 5983 labelled images. To further compare with the state-of-the-art supervised methods, we also conducted experiments using the whole 5983 images of the training set as a labelled set, and used 17,617 additional unlabelled frames from the original videos.

For EndoVis 2017, as no additional unlabeled data is available, the original training set, which has 1800 images in total, is split into labelled and unlabelled subsets with ratios 1:9 or 3:7.

As for the multi-class segmentation evaluation using CholecInstanceSeg, we use the same data splitting ratios as in our experiments with the Robust-MIS 2019 dataset. In our ablation studies, we utilized only the original 26.8k labelled training images, retaining 30% of the annotations. To further evaluate multi-class segmentation performance and compare it against the baseline model, we also used the full set of 26.8k labelled training images and the 3804 frames in the validation set as the labelled dataset. Additionally, we incorporated a random selection of 66.1k unlabelled frames to assess the impact of including unlabelled data. For testing, we utilized the 11,299 frames from the original test split.

Implementation details
During training, for each batch, the same number of images is sampled from the labelled dataset D_l and the unlabelled dataset D_u. Within each batch, unlabelled samples are initialized by random but hand-crafted strong augmentations and then adversarially updated by adding the I-FGSM perturbations.

All our experiments were trained on two NVIDIA V100 (32GB) with a learning rate of 0.001. The model was trained using an SGD optimizer with a momentum of 0.95, which we found can provide a smooth convergence trajectory (unreported manual tuning involved experiments with momentum values from 0.9 to 0.99). The learning rate was initialized as 1.0×10^{-2} and decayed with the policy $lr_{ini} \times (1 - epoch/epoch_{max}) \times \eta$, where $epoch_{max}$ is the total epochs and η is set to 0.7. The total training epoch number is 1000 and the batch size was set as 64 considering the memory constraints and training efficiency. The selection of these parameters was initially based on the baseline model OR-Unet[7] and further adjusted through multiple experimental runs and prior experience with similar setups.

For the hyper-parameters in adversarial augmentation, the magnitude of adversarial augmentations ϵ was set to 0.08 based on experiments shown in Fig. 6. This value represents the optimal trade-off, which demonstrates that increasing ϵ beyond 0.08 led to performance degradation. As shown in Table 7 and Fig. 7 in the manuscript, increasing I-FGSM iterations improved segmentation performance slightly. However, considering the trade-off between performance and computational efficiency, we selected step number $K = 25$ in our final model. Additionally, two types of initial strong augmentation techniques were utilized with pre-defined minimum and maximum magnitude values, specifically from the photometric transformations in RandAugment[48] method.

For the segmentation model backbone, we employed the representative OR-Unet[7]. Our OR-Unet contains 6 downsampling and 6 upsampling blocks where the residual blocks were employed in the encoder and the sequence Conv-BN-Relu layers with kernel size 3×3 were used in the decoder.

Evaluation metrics
We evaluated our model based on the criterion proposed in Robust-MIS 2019 MICCAI challenge[1] for binary segmentation task, which includes:

- Dice Similarity Coefficient, a widely used overlap metric in segmentation challenges;

- Normalized Surface Dice (NSD)[1] is a distance-based measurement for assessing performance, which measures the overlap of two surfaces (i.e. mask borders). In adherence to the challenge guidelines[1], we set the tolerance value for NSD to 13 pixels, which takes into account inter-rater variability. Note that the value of tolerance was determined by comparing annotations from five annotators on 100 training images as illustrated in the challenge.

For the multi-class segmentation evaluation, we adopted the evaluation method from ISINet[31] and used three IoU-based metrics: Ch_IoU, ISI_IoU, and mc_IoU.

- Ch_IoU calculates the mean IoU for each category present in the ground truth of an image, then averages these values across all images.
- ISI_IoU extends Ch_IoU by computing the mean IoU for all predicted categories, regardless of whether they are present in the ground truth of the image. Typically, Ch_IoU is greater than or equal to ISI_IoU.
- mc_IoU calculates the average IoU across different instrument classes and addresses category imbalance by altering the averaging order used in ISI_IoU.

We used "percentage points" (pp) to denote the absolute change in Mean Dice score, NSD, Ch_IoU, ISI_IoU, and mc_IoU, as it represents the difference between two percentage values. Additionally, as shown in Sect. Comparison with strong baselines, we report both the mean performance and the inter-run standard deviation (std) across folds for our proposed model, SegMatch, to capture performance consistency across multiple runs. For other models, we provide the single-run sample-based std, which reflects consistency within a single run. In Seciton Ablation and parameter sensitivity study, we report the single-run sample-based std for both the original SegMatch model and the other ablation experiments.

Results

Comparison with strong baselines

We compare our results to the state-of-the-art on Robust-MIS 2019 and EndoVis 2017 datasets. We categorize comparisons into two groups. First, a head-to-head comparison is made with other semi-supervised methods. Second, we measure the added value of incorporating unlabelled data in addition to using the complete labelled training data in Robust-MIS 2019.

Comparison with semi-supervised baselines

For the first group, we adapted the representative semi-supervised classification method Mean-Teacher[19] for the segmentation task using the same backbone network and experimental setting as ours. We also conducted experiments with two established semi-supervised semantic segmentation models: WSSL[55], a representative benchmark in this field, and CCT[27], which emphasizes feature consistency across various contexts through perturbed feature alignment. Additionally, we evaluated ClassMix[56], a novel data augmentation technique designed for semantic segmentation, Min-Max Similarity[57], a semi-supervised surgical instrument segmentation network based on contrastive learning, and PseudoSeg[58], a FixMatch-based semi-supervised semantic segmentation model with a self-attention module. Illustrative segmentation results for the selective methods on Robust-MIS 2019 are presented in Fig. 4, with key areas highlighted for clarity.

Table 1 shows that, within the dataset of Robust-MIS 2019, for the two labelled to unlabelled data ratios we tested, our SegMatch outperforms other methods with statistical significance (p<0.05). Comparing SegMatch to the second-best method, PseudoSeg, we observed notable performance improvements. Specifically, when using

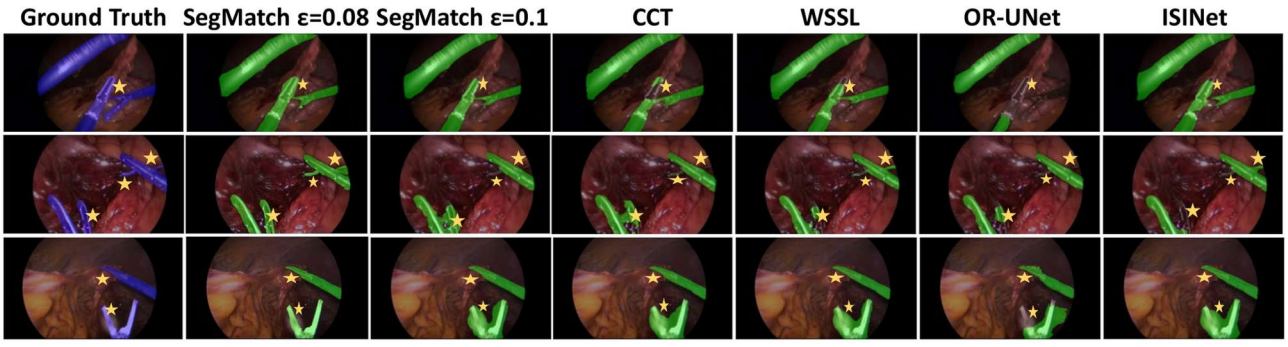


Fig. 4. Segmentation results on exemplar images from three different procedures in the testing set. Here, SegMatch, CCT[27], and WSSL[55] were trained using the whole labelled training set of Robust-MIS 2019 as a labelled set, and 17K additional unlabelled frames from the original videos. The fully supervised learning models (OR-UNet[7] and ISINet[31]) were trained using the whole labelled training set of Robust-MIS 2019 as a labelled set. The first column is the ground truth mask placed on top of the original image, and the other column is the segmentation results of SegMatch ablation models and state-of-the-art models. The three rows from up to bottom are the testing image samples from proctocolectomy procedures, sigmoid resection procedure (unseen type), and rectal resection procedure respectively. The yellow stars highlight the key area in better visualization..

Methods	Robust-MIS 2019 (5983 training images)				EndoVis 2017 (5286 training images)			
	Labelled: 598 (10%)		Labelled: 1794 (30%)		Labelled: 528 (10%)		Labelled:1585 (30%)	
	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD
Mean-teacher[19]	62.1± 3.33	61.8 ± 3.39	70.2 ± 4.47	69.0 ± 4.45	51.6 ± 3.27	52.7 ± 3.22	60.2 ± 4.28	59.9 ± 4.26
WSSL[55]	64.3 ± 1.56	72.1 ± 2.58	59.2 ± 2.64	67.5 ± 2.71	62.6 ± 1.85	73.2 ± 2.32	60.4 ± 3.1	70.2 ± 2.77
CCT[27]	67.1 ± 2.3	65.2 ± 2.6	75.2 ± 3.1	72.6 ± 2.1	61.2 ± 1.9	58.7 ± 2.8	69.6 ± 2.4	65.5 ± 2.0
ClassMix[56]	67.5 ± 2.1	61.9 ± 1.8	73.2 ± 3.0	70.6 ± 2.7	74.7 ± 2.5	74.3 ± 2.2	78.6 ± 3.1	72.9 ± 2.5
Min-Max Similarity[57]	68.7 ± 3.13	61.2 ± 2.47	78.6 ± 3.29	72.1 ± 2.94	79.4 ± 3.18	77.9 ± 2.61	87.1 ± 3.36	76.4 ± 3.05
PseudoSeg[58]	70.4 ± 2.78	67.5 ± 3.01	79.9 ± 2.92	73.4 ± 2.64	79.8 ± 3.11	78.7 ± 2.85	87.8 ± 3.23	75.5 ± 2.99
SegMatch	73.3 ± 1.8	71.2 ± 1.4	84.3 ± 2.3	80.2 ± 1.9	81.1 ± 2.1	76.3 ± 1.7	89.2 ± 2.5	72.0 ± 1.6
↪ std across folds	± 0.42	± 0.48	± 0.61	± 0.51	± 0.26	± 0.42	± 0.33	± 0.45

Table 1. State-of-the-art semi-supervised model comparisons for Robust-MIS 2019 dataset (left) and EndoVis 2017 dataset (right) under differently labelled data to unlabelled data ratio. The reported standard deviation is measured from the performance across the testing data in a single run, except for SegMatch where we also report the standard deviation of the mean score across different runs.

Labelled Data Ratio										
	598 (10%)		30%		50%		70%		100%	
	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD
OR-Unet [7]	62.1 ± 2.9	60.4 ± 2.62	72.6 ± 1.76	71.2 ± 2.34	75.2 ± 2.69	73.6 ± 2.28	80.1 ± 1.45	77.5 ± 1.72	88.0 ± 1.22	86.2 ± 1.07
SegMatch	73.3 ± 1.89	71.2 ± 1.4	76.3 ± 2.3	74.2 ± 1.9	80.9 ± 2.65	78.6 ± 1.74	82.7 ± 1.42	79.2 ± 2.15	87.5 ± 1.09	84.9 ± 0.98

Table 2. Performance comparison for OR-Unet[7] (fully supervised methods) vs SegMatch (our semi-supervised model) across different labelled data ratios on Robust-MIS 2019 dataset.

Method	Labelled: 5983 training images (+17K unlabelled if semi-supervised)							
	Whole testing		Stage 1		Stage 2		Stage 3	
	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD	Mean Dice	NSD
OR-Unet[7]*	88.0 ± 1.22	86.2 ± 1.07	90.2 ± 1.15	88.5 ± 1.3	87.9 ± 1.1	85.6 ± 1.05	85.9 ± 1.2	84.5 ± 1.18
Robust-MIS 2019 winner[1]*	90.1 ± 1.15	88.9 ± 1.08	92.0 ± 1.25	92.7 ± 1.1	90.2 ± 1.18	88.6 ± 1.05	89.0 ± 1.11	86.4 ± 1.09
ISINet[31]*	88.9 ± 1.14	86.3 ± 1.07	90.9 ± 1.22	87.6 ± 1.15	89.6 ± 1.1	86.5 ± 1.08	86.2 ± 1.05	84.0 ± 1.12
DINO-Adapter[36]*	92.9 ± 1.01	91.5 ± 0.97	94.2 ± 1.08	92.4 ± 1.12	92.6 ± 0.94	91.4 ± 1.17	91.9 ± 1.09	90.7 ± 0.93
SegMatch	93.7 ± 1.02	93.6 ± 0.99	95.1 ± 1.25	95.5 ± 1.18	93.1 ± 1.35	92.5 ± 1.07	92.9 ± 1.27	92.8 ± 0.98
↪ std across folds	± 0.28	± 0.24	± 0.23	± 0.19	± 0.49	± 0.31	± 0.12	± 0.22

Table 3. Comparison on the Robust-MIS 2019 dataset between fully-supervised models and SegMatch with additional unlabelled data (* indicates only labelled data was used). The reported standard deviation is measured from the performance across the testing data in a single run, expect for SegMatch where we also report the standard deviation of the mean score across different runs.

10% and 30% labelled data in the Robust-MIS 2019 dataset, SegMatch achieved mean Dice score improvements of 2.9 percentage points (pp) and 4.4 pp, respectively. Similar observations were made on the EndoVis 2017 dataset, where SegMatch outperformed PseudoSeg by 1.3 pp and 1.4 pp when utilizing 10% and 30% labelled data, respectively.

These results demonstrate the superior performance of SegMatch in both datasets and both labelled to unlabelled data ratios, highlighting the effectiveness of our proposed method in scenarios with limited labelled data. Note that, the training dataset only consists of the proctocolectomy procedure and rectal resection procedure, and the sigmoid resection procedure is considered a new type of data for the trained model. Qualitatively, our proposed SegMatch is able to recognize the boundaries between different tools and segment more complete shapes for each individual tool, especially in those areas with high reflection.

Comparison with fully-supervised method

We compare OR-Unet[7] with our SegMatch across a wide range of labelled-to-unlabelled ratios of the Robust-MIS 2019. We train the fully supervised OR-Unet[7] and our semi-supervised SegMatch using the same amount of labelled data while utilising the remaining data as unlabelled data only for SegMatch.

The results shown in Table 2 demonstrate that SegMatch outperforms OR-Unet consistently by effectively utilizing additional unlabelled data. As the amount of labelled data increases, the benefit of SegMatch diminishes,

leading to the convergence of the performance of the two methods. Nonetheless, minor variations in the 100% labelled case arose due to differences in codebase implementation and experimental randomness.

Added value of unlabelled data

For the second group, we used the whole labelled training set of Robust-MIS 2019 as the labelled set and take advantage of the unlabelled video footage available from the training set of Robust-MIS 2019 to evaluate the impact of adding unlabelled data as mentioned in Sect. Dataset.

Table 3 shows the comparison between supervised approaches trained on the labelled set and SegMatch trained on the combination of the labelled and unlabelled sets. Compared to the existing model OR-Unet[7], the inclusion of additional unlabelled data in our semi-supervised pipeline SegMatch led to a 5.7 pp higher Dice score. Our model also demonstrates a noteworthy enhancement of 4.8 pp in comparison to the more recent ISInet[31], which is now commonly employed for surgical instrument segmentation. Additionally, it maintains a 0.8 pp advantage compared to DINO-Adapter[36], a recent approach that leverages pre-trained knowledge from foundation models via adapter networks When evaluating the stage 3 testing data, which corresponds to a surgical procedure that was not seen during the training phase, our SegMatch model demonstrated superior performance compared to the official Robust-MIS 2019 challenge winner for stage 3 (haoyun team) by a margin of 3.9 pp.

It is noteworthy that the performance improvement from SegMatch over fully-supervised baselines is more substantial on stage 3 testing data (unseen procedure types) than on stage 1 and 2 data (procedure types represented in the training data). This indicates that our model exhibits better generalizability. This enhanced generalizability can enable the model to handle diverse surgical scenarios more effectively.

Multi-class segmentation

To evaluate the multi-class segmentation capabilities of our model, we conducted experiments using the CholecInstanceSeg dataset. We utilized the entire labelled training set and leveraged the available unlabeled video footage to analyze the impact of incorporating unlabelled data, as discussed in Sec. Dataset.

Table 4 compares our method with the baseline OR-Unet approach[7]. By incorporating an additional 66.1k unlabeled data points, our model achieved significant improvements, surpassing OR-Unet by 24.2 in Ch_IoU, 21.89 in ISI_IoU, and 25.06 in mc_IoU. Additionally, our approach outperforms OR-Unet across multiple instrument categories, highlighting the effectiveness of utilizing unlabeled data in enhancing multi-class segmentation compared to the fully-supervised OR-Unet.

Ablation and parameter sensitivity study

To evaluate the contribution of the different components of our pipeline, we conducted ablation and parameter sensitivity studies on different strong augmentation strategies and adversarial strong augmentation methods.

Analysis of our semi-supervised method

We evaluated the contribution of the semi-supervised learning in SegMatch by training only its fully-supervised branch, essentially turning it into OR-UNet. As discussed previously and shown in Table 3, disabling the unlabelled pathway leads to a drop in terms of Dice and NSD scores thereby confirming the benefits of semi-supervised learning.

We also studied the effectiveness of varying the confidence threshold when generating the pseudo-labels as shown in Fig. 5. We observe that when the confidence threshold approaches 1.0, the model returns the worst segmentation performance. When the threshold value is changed within the range of [0.7, 0.9], the confidence threshold does not affect the model's performance significantly. However, it should be noted that further reducing the threshold value leads to a rapid decrease in the mean Dice score, which indicates that the quality of contributing pixels to the unsupervised loss may be more important than the quantity. This evaluation was conducted on the binary segmentation task using the Robust-MIS 2019 dataset, but a similar trend was observed in the multi-class segmentation task as well.

Augmentation strategy

We also operated an ablation study on the weak and strong augmentation in SegMatch as tabulated in Table 5. First, we removed the adversarial augmentation, thereby only keeping handcrafted strong augmentations. Notably, we observed consistent results across different labelled data ratios. Specifically, when utilizing 17K additional unlabelled data, we found that the Dice score decreased by 3.1 pp compared to the full proposed SegMatch

	Labelled: 26.8K training images (66.1K unlabelled if semi-supervised)									
Method	**Ch_IoU**	**ISI_IoU**	**Grasper**	**Bipolar**	**Hook**	**Clipper**	**Scissors**	**Irrigator**	**Snare**	**mc_IoU**
OR-Unet*[7]	65.62 ± 1.23	62.20 ± 1.45	48.70 ± 1.58	48.50 ± 1.36	60.09 ± 1.72	37.43 ± 1.42	12.01 ± 1.11	38.72 ± 1.67	22.56 ± 1.29	38.96 ± 1.54
SegMatch	89.82 ± 1.43	84.09 ± 1.28	80.12 ± 1.67	79.74 ± 1.15	72.29 ± 1.53	46.96 ± 1.21	32.92 ± 1.05	72.12 ± 1.62	63.96 ± 1.47	64.02 ± 1.38
↪ std across folds	± 0.31	± 0.32	± 0.42	± 0.37	± 0.52	± 0.27	± 0.34	± 0.43	± 0.26	± 0.29

Table 4. Comparison of our method with state-of-the-art methods on the CholecInstanceSeg dataset for multi-class segmentation. (* indicates only labelled data was used). The reported standard deviation is measured from the performance across the testing data in a single run, expect for SegMatch where we also report the standard deviation of the mean score across different runs.

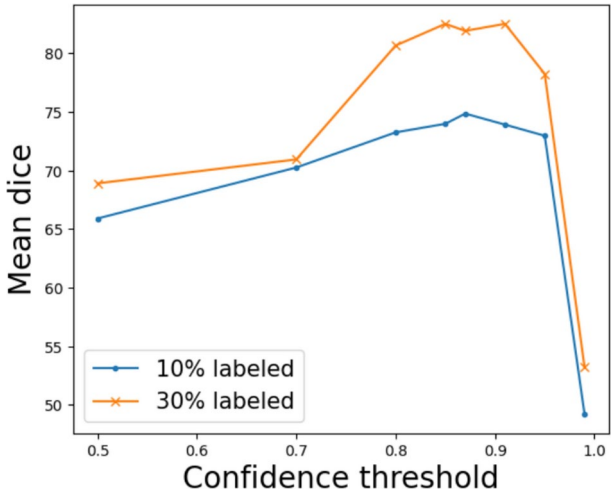


Fig. 5. Mean Dice score produced by varying the confidence threshold for pseudo-labels.

Method	Labelled: 1794 (30%)		Labelled: 5983 (+17K unlabelled)	
	Mean Dice	NSD	Mean dice	NSD
SegMatch	84.3 ± 2.3	80.2 ± 1.9	93.7 ± 1.02	93.6 ± 0.99
No adversarial	80.6 ± 1.67	77.9 ± 1.74	90.6 ± 1.87	89.9 ± 1.96
No adversarial & No weak aug	75.6 ± 1.62	73.7 ± 1.76	88.7 ± 1.59	87.6 ± 1.69
No adversarial & No strong aug	71.5 ± 1.45	68.2 ± 1.37	85.5 ± 1.39	86.9 ± 1.42

Table 5. Ablation study results for different components, evaluating on Robust-MIS 2019 with different labelled and unlabelled data amount, on binary segmentation.

Method	Labelled: 8.04K (30%)		Labelled: 26.8K (+66.1K unlabelled)	
	Ch_IoU	ISI_IoU	Ch_IoU	ISI_IoU
SegMatch	72.12 ± 2.49	69.04 ± 2.52	89.92 ± 1.43	84.09 ± 1.28
No adversarial	67.34 ± 1.97	61.27 ± 1.92	84.25 ± 1.86	81.29 ± 1.78
No adversarial & No weak aug	55.24 ± 1.92	53.27 ± 1.99	80.26 ± 1.79	77.06 ± 1.72
No adversarial & No strong aug	54.32 ± 1.66	52.66 ± 1.64	76.90 ± 1.77	72.34 ± 1.75

Table 6. Ablation study results for different components, evaluating on CholecInstanceSeg with different labelled and unlabelled data amount, on multi-class segmentation.

model. This suggests that applying adversarial augmentations can prevent the learning saturation caused by hand-crafted strong augmentations on unlabelled data, thereby enabling the model to learn continuously. When both the weak augmentation and adversarial augmentation are removed, the results drop by an additional 1.9 pp compared to only removing the adversarial augmentation, indicating that applying the weak augmentation function to the input image, which generates new and diverse examples for the training dataset, can enhance the segmentation performance.

To evaluate the effectiveness of the overall strong augmentation strategies, we replaced the strongly augmented images with the original images. The resulting Dice score dropped by 5.1 pp compared to only removing the adversarial augmentation, and by 3.2 pp compared to removing both the weak and adversarial augmentation. The evidence that removing strong augmentation results in a greater decrease in performance than removing weak augmentation suggests that stronger perturbations are beneficial for learning with consistency regularization. Additionally, when strong augmentation is removed, the model is presented with the same input image from different views, which reduces the benefits of consistency regularization. In this scenario, pseudo-labelling becomes the primary technique for achieving better segmentation performance. Therefore, our results suggest that consistency regularization is more crucial than pseudo-labelling in our pipeline for improving segmentation performance.

We conducted an ablation study on augmentation strategies for multi-class segmentation using the CholecInstanceSeg dataset, as shown in Table 6. Even when trained on only 70% of the data, our model

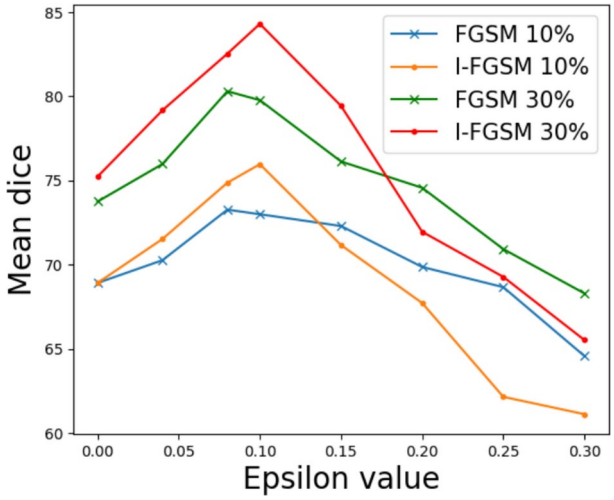


Fig. 6. Optimal ϵ value enhances segmentation performance (as indicated by the peak in mean Dice score).

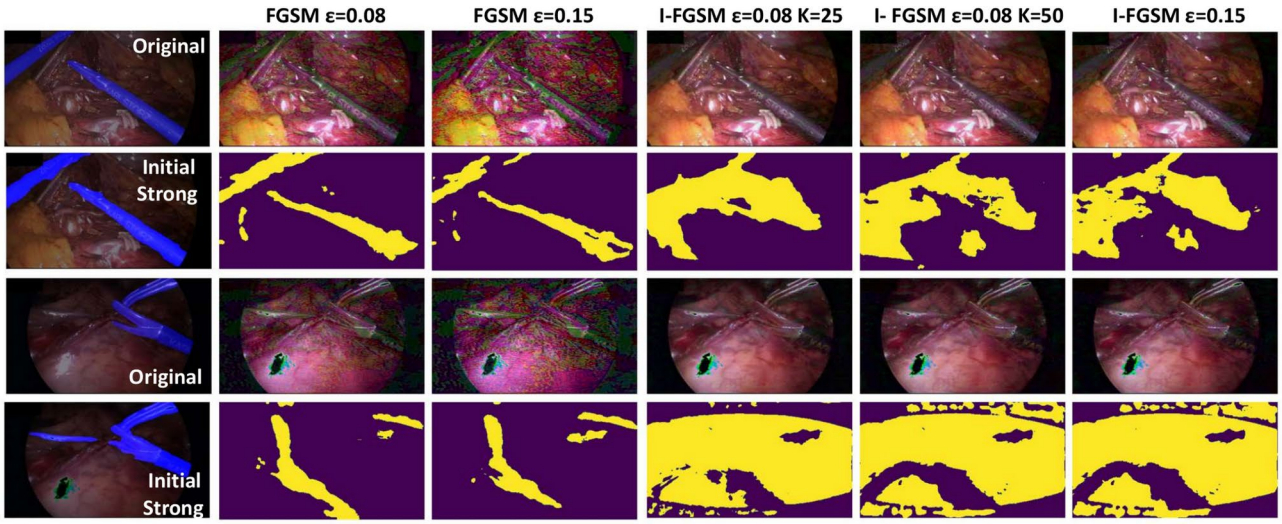


Fig. 7. Examples showcase the impact of strong augmentation transform functions and adversarial augmentation on an original unlabelled image input to a model. The first column features the original image covered by its segmentation mask output from the model, as well as the strongly augmented image obtained via initial strong augmentation and its output segmentation mask. The 2-6 columns showcase adversarial images produced by I-FGSM with varying values of ϵ and K (which becomes FGSM when $K = 0$) to replace the original strongly augmented images for model parameter updating. The upper rows in the 2-6 columns display the adversarial images, while the bottom rows show the corresponding segmentation results produced by the model.

outperformed the baseline OR-Unet[7], which was trained on the full dataset as shown in Table 4, by 6.5 pp in Ch_IoU and 6.84 pp in ISI_IoU. This demonstrates the potential of our model in scenarios with limited annotated data. Consistent with results from the binary segmentation task, removing adversarial augmentation led to a noticeable drop in IoU scores, highlighting its role in preventing learning saturation from over-reliance on hand-crafted strong augmentations in unlabelled data. The results show that removing strong augmentation hurts performance more than removing weak augmentation, suggesting that, similar to binary segmentation, multi-class segmentation benefits from a greater input difference between branches for effective consistency regularization.

Adversarial augmentation analysis

We further evaluated the sensitivity of our results to changes in the maximum amplitude value ϵ of the adversarial perturbation as shown in Fig. 6 and qualitatively illustrated in Fig. 7. When increasing from $\epsilon = 0.0$ (i.e. no perturbation), we observed a consistent pattern across different ratios of labelled data for both FGSM and I-FGSM. Initially, as ϵ increased, the segmentation performance improved, reaching its peak at approximately

Adversarial method (within SegMatch)	Labelled: 1794(30%)	
	Mean Dice	NSD
C&W[43]	72.9 ± 2.22	70.4 ± 2.13
FGSM[41]	75.8 ± 2.16	79.6 ± 2.05
DDN[44]	78.9 ± 2.21	77.2 ± 2.52
I-FGSM ($\epsilon = 0.08, K = 25$)	84.3 ± 2.3	80.2 ± 1.9
I-FGSM ($\epsilon = 0.08, K = 50$)	84.9 ± 2.19	81.5 ± 2.02

Table 7. Segmentation performance of SegMatch on Robust-MIS 2019 when applying different adversarial attack methods, evaluating on binary segmentation task.

Adversarial method (within SegMatch)	Labelled: 12.57K (30%)	
	Ch_IoU	ISI_IoU
C&W[43]	64.45 ± 2.58	62.12 ± 2.61
FGSM[41]	65.76 ± 2.39	64.39 ± 2.42
DDN[44]	67.24 ± 2.68	66.56 ± 2.75
I-FGSM ($\epsilon = 0.08, K = 25$)	72.12 ± 2.49	69.04 ± 2.52
I-FGSM ($\epsilon = 0.08, K = 50$)	74.32 ± 2.28	71.29 ± 2.31

Table 8. Segmentation performance of SegMatch on CholecInstanceSeg when applying different adversarial attack methods, evaluating on multi-class segmentation task.

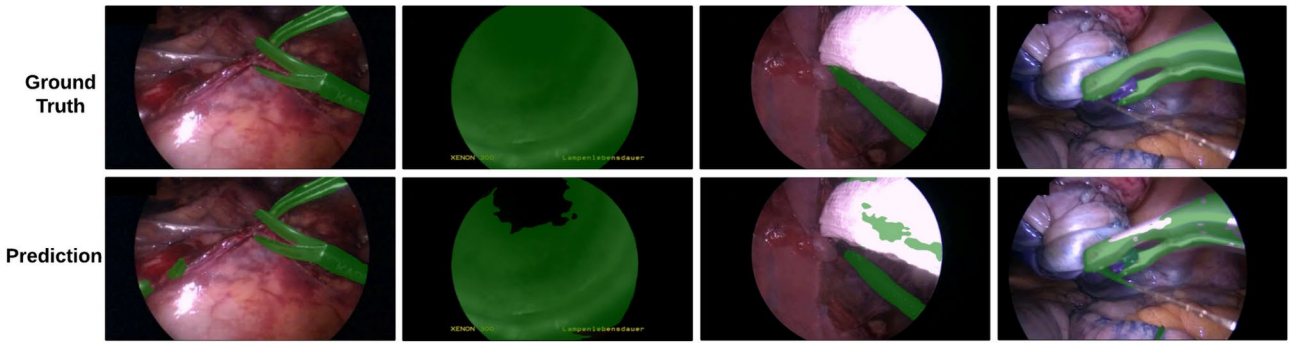


Fig. 8. Failure cases of SegMatch's output. The first row shows the original image with the ground truth mask, while the second row overlays SegMatch predictions on the original image. Columns 1 to 4 illustrate the 1st, 2nd, 3rd, and 4th examples, respectively.

$\epsilon = 0.08$. However, beyond this optimal point, the performance started to decline, indicating that stronger perturbations can enhance the model's performance only within a certain range.

In this work, by precisely defining the acceptable range of perturbations, we restrict the perturbations within the ϵ-neighbourhood, which ensures that the integrity of the instrument pixels is preserved while introducing subtle variations that aid in improved generalization of the model. Comparing FGSM and I-FGSM, we found that I-FGSM showed superior performance to FGSM before reaching the optimal ϵ value. However, after the optimal point, I-FGSM exhibited a more significant decrease in the model's performance compared to FGSM, which suggests that I-FGSM has a higher attack success rate than FGSM but becomes more harmful to the model when the attacking amplitude is large.

We also varied the number of I-FGSM iterations as shown in Table 7 and qualitatively illustrated in Fig. 7. Increasing the number of iterations in the I-FGSM attack shows a minor improvement consistent with expectations that it increases the attach success rate. However, it is important to consider that this small improvement in segmentation performance through increased iterations comes with a trade-off in computational efficiency.

Additionally, we conducted experiments using various adversarial learning methods[43,44] as shown in Table 7. Our findings indicate that performance achieved by using one-step attack methods is consistently lower compared to our iterative strategy, which suggests when attempting to manipulate image samples to create adversarial examples, breaking the attack down into smaller steps can improve the overall success rate, as noted in previous research[14].

The same ablation studies for adversarial attack methods were conducted on multi-class segmentation tasks using the CholecInstanceSeg dataset, as summarized in Table 8. The results consistently align with

those observed in binary segmentation, demonstrating that one-step attack approaches perform worse than our interactive adversarial augmentation. Notably, when the maximum perturbation amplitude ϵ is fixed, the number of iterations has a greater impact on the performance of multi-class segmentation compared to binary segmentation.

Failure cases analysis

We present failure cases in Fig. 8, highlighting challenges where SegMatch struggles to accurately segment surgical instruments. While adversarial and trainable augmentations allow flexible transformations to handle complex scenarios like overlapping tools or instruments obscured by tissue, the model faces difficulties with reflective surfaces (e.g., gauze in the 3rd example) and instruments resembling the tissue background (4th example), leading to false positives and imprecise pseudo-labels.

This implies noise in unlabelled data, such as poor image quality or artefacts, can reduce SegMatch's performance since consistency regularization may amplify errors. Additionally, out-of-distribution scenarios, such as the instrument-dominated image in the 2nd example, further challenge SegMatch, demonstrating its limitations in adapting to highly variable or ambiguous inputs despite leveraging unlabelled data effectively.

Discussion and conclusion

In this paper, we introduced SegMatch, a semi-supervised learning algorithm for surgical tool segmentation that achieves state-of-the-art results across two of the most commonly used datasets in this field. Our algorithm, SegMatch, was adapted from a simple semi-supervised classification algorithm FixMatch which combines consistency regularization and pseudo-labelling. During training, SegMatch makes use of a standard labelled image pathway and an unlabelled image pathway with training batches mixing labelled and unlabelled images. The unlabelled image pathway is composed of two concurrent branches. A weak augmentation branch is used to generate pseudo-labels against which the output of a strong augmentation branch is compared. Considering the limitation of fixed handcrafted strongly augmentation techniques, we introduced adversarial augmentations to increase the performance of strongly augmented images. We also highlighted the importance of considering equivariance and invariance properties in the augmentation functions used for segmentation.

Putting our work into its application context, automatic visual understanding is critical for advanced surgical assistance but labelled data to train supporting algorithms are expensive and difficult to obtain.

SegMatch's ability to leverage unlabelled data makes it highly applicable to clinical environments, where appropriately annotated data is severely restricted due to time and budget constraints. Moreover, SegMatch's adaptability to varying levels of supervision aligns well with clinical workflows, where relying on partial annotations or weak labels is more feasible than the resource-intensive process of creating fully annotated datasets.

In robotic-assisted surgeries, our approach enables precise segmentation of surgical instruments during intraoperative procedures, facilitating real-time image analysis and delivering enhanced visual guidance to support surgical decision-making. Key applications include augmented reality for precise overlay placement, haptic feedback simulation to improve maneuver safety, and tissue mosaicking to broaden the surgical visual field. Furthermore, these advancements can contribute to surgical workflow optimization, objective skills assessment, camera calibration, and visual servoing, paving the way for improved accuracy, efficiency, and automation in complex surgical interventions.

The principles underlying SegMatch are not confined to surgical instrument segmentation. Our trainable adversarial augmentation strategy could be equally beneficial in domains requiring precise segmentation of complex structures whereas annotations are limited and expensive to obtain.

We believe that simple but strong-performance semi-supervised segmentation learning algorithms, such as our proposed SegMatch, will not only accelerate the deployment of surgical instrument segmentation in the operating theatre but also be applicable to other domains lacking annotated data.

Data availability

The datasets used in this study are publicly available as follows: MICCAI Instrument Segmentation Challenge datasets: Robust-MIS 2019 - accessible at https://www.synapse.org/Synapse:syn20575265. EndoVis 2017 - acc essible at https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/. The CholecInstanceSe g dataset: Available at https://www.synapse.org/Synapse:syn60239970/datasets/. These datasets are distributed under their respective open-access terms and do not require additional permissions for use.

Received: 11 November 2024; Accepted: 14 March 2025
Published online: 23 April 2025

References

1. Ross, T. et al. Robust medical instrument segmentation challenge 2019. *arXiv preprint* arXiv:2003.10299 (2020).
2. Speidel, S. et al. Visual tracking of da Vinci instruments for laparoscopic surgery. In: *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 9036, 47–52 (SPIE, 2014).
3. Rieke, N. et al. Real-time localization of articulated surgical instruments in retinal microsurgery. *Med. Image Anal.* **34**, 82–100 (2016).
4. Laina, I. et al. Concurrent segmentation and localization for tracking of surgical instruments. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 664–672 (Springer, 2017).
5. Hasan, S. K. & Linte, C. A. U-NetPlus: A modified encoder-decoder U-Net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 7205–7211 (IEEE, 2019).

6. Islam, M., Li, Y. & Ren, H. Learning where to look while tracking instruments in robot-assisted surgery. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 412–420 (Springer, 2019).
7. Isensee, F. & Maier-Hein, K. H. OR-UNet: an optimized robust residual U-Net for instrument segmentation in endoscopic images. *arXiv preprint* arXiv:2004.12668 (2020).
8. Sahu, M., Mukhopadhyay, A. & Zachow, S. Simulation-to-real domain adaptation with teacher-student learning for endoscopic instrument segmentation. *Int. J. Comput. Assist. Radiol. Surg.* **16**, 849–859 (2021).
9. Ross, T. et al. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. *Int. J. Comput. Assist. Radiol. Surg.* **13**, 925–933 (2018).
10. Liu, D. et al. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 657–667 (Springer, 2020).
11. Sanchez-Matilla, R., Robu, M., Luengo, I. & Stoyanov, D. Scalable joint detection and segmentation of surgical instruments with weak supervision. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 501–511 (Springer, 2021).
12. Zhao, Z., Jin, Y., Gao, X., Dou, Q. & Heng, P.-A. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 679–689 (Springer, 2020).
13. Sohn, K. et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inform. Process. Syst.* **33**, 596–608 (2020).
14. Kurakin, A., Goodfellow, I. J. & Bengio, S. Adversarial examples in the physical world. In: *Artificial Intelligence Safety and Security*, 99–112 (Chapman and Hall/CRC, 2018).
15. Allan, M. et al. 2017 robotic instrument segmentation challenge. *arXiv preprint* arXiv:1902.06426 (2019).
16. Alabi, O. et al. Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery (2024). arXiv:2406.16039.
17. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. Self-training with noisy student improves imagenet classification. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10687–10698 (2020).
18. Qiao, S., Shen, W., Zhang, Z., Wang, B. & Yuille, A. Deep co-training for semi-supervised image recognition. In: *Proceedings of the European Conference on Computer Vision*, 135–152 (2018).
19. Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: *Advances in Neural Information Processing Systems*, 30 (2017).
20. Rasmus, A., Berglund, M., Honkala, M., Valpola, H. & Raiko, T. Semi-supervised learning with ladder networks. In: *Advances in Neural Information Processing Systems*, 28 (2015).
21. Berthelot, D. *et al.* MixMatch: A holistic approach to semi-supervised learning. In: *Advances in Neural Information Processing Systems*, 32 (2019).
22. Grandvalet, Y. & Bengio, Y. Semi-supervised learning by entropy minimization. In: *Advances in Neural Information Processing Systems*, 17 (2004).
23. Kim, J., Jang, J. & Park, H. Structured consistency loss for semi-supervised semantic segmentation. *arXiv preprint* arXiv:2001.04647 (2020).
24. Li, X., Yu, L., Chen, H., Fu, C.-W. & Heng, P.-A. Semi-supervised skin lesion segmentation via transformation consistent self-ensembling model. In: *British Machine Vision Conference* (2018).
25. Feng, Z. *et al.* DMT: Dynamic mutual training for semi-supervised learning. *Pattern Recognition* 108777 (2022).
26. Chen, L.-C. *et al.* Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In: *Proceedings of the European Conference on Computer Vision*, 695–714 (Springer, 2020).
27. Ouali, Y., Hudelot, C. & Tami, M. Semi-supervised semantic segmentation with cross-consistency training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684 (2020).
28. Chen, X., Yuan, Y., Zeng, G. & Wang, J. Semi-supervised semantic segmentation with cross pseudo supervision. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2613–2622 (2021).
29. Souly, N., Spampinato, C. & Shah, M. Semi supervised semantic segmentation using generative adversarial network. In: *Proceedings of the IEEE International Conference on Computer Vision*, 5688–5696 (2017).
30. Wang, Y., Xuan, Z., Ho, C. & Qi, G.-J. Adversarial dense contrastive learning for semi-supervised semantic segmentation. in: *IEEE Transactions on Image Processing* (2023).
31. González, C., Bravo-Sánchez, L. & Arbelaez, P. Isinet: an instance-based approach for surgical instrument segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 595–605 (Springer, 2020).
32. Qin, F., Li, Y., Su, Y.-H., Xu, D. & Hannaford, B. Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose. In: *2019 International Conference on Robotics and Automation*, 9821–9827 (IEEE, 2019).
33. Jin, Y., Cheng, K., Dou, Q. & Heng, P.-A. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 440–448 (Springer, 2019).
34. Seenivasan, L., Mitheran, S., Islam, M. & Ren, H. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robot. Automa. Lett.* **7**, 3858–3865 (2022).
35. Zhao, Z., Jin, Y. & Heng, P.-A. Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery. In: *ICRA*, 11186–11193 (IEEE, 2022).
36. Wei, M., Shi, M. & Vercauteren, T. Enhancing surgical instrument segmentation: integrating vision transformer insights with adapter. *International Journal of Computer Assisted Radiology and Surgery* 1–8 (2024).
37. Liu, J., Guo, X. & Yuan, Y. Graph-based surgical instrument adaptive segmentation via domain-common knowledge. *IEEE Trans. Med. Imaging* **41**, 715–726 (2021).
38. Goodfellow, I. *et al.* Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, 27 (2014).
39. Hinton, G., Vinyals, O., Dean, J. et al. Distilling the knowledge in a neural network. *arXiv preprint* arXiv:1503.02531**2** (2015).
40. Antoniou, A., Storkey, A. & Edwards, H. Data augmentation generative adversarial networks. *arXiv preprint* arXiv:1711.04340 (2017).
41. Szegedy, C. et al. Intriguing properties of neural networks. *arXiv preprint* arXiv:1312.6199 (2013).
42. Tramèr, F. et al. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations* (2018).
43. Carlini, N. & Wagner, D. Towards evaluating the robustness of neural networks. In: *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57 (IEEE, 2017).
44. Rony, J. et al. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4322–4330 (2019).
45. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In: *International Conference on Learning Representations* (2018).
46. Samangouei, P., Kabkab, M. & Chellappa, R. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In: *International Conference on Learning Representations* (2018).
47. Cohen, J., Rosenfeld, E. & Kolter, Z. Certified adversarial robustness via randomized smoothing. In: *International Conference on Machine Learning*, 1310–1320 (PMLR, 2019).

48. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. RandAugment: Practical automated data augmentation with a reduced search space. In: *Proceedings of the IEEE/CVF Conference on Computer Vvision and Pattern Recognition Workshops*, 702–703 (2020).
49. Berthelot, D. *et al.* ReMixMatch: Semi-supervised learning with distribution matching and augmentation anchoring. In: *International Conference on Learning Representations* (2020).
50. Garcia-Peraza-Herrera, L. C. et al. Image compositing for segmentation of surgical tools without manual annotations. *IEEE Trans. Med. Imaging* **40**, 1450–1460 (2021).
51. Minaee, S. et al. Image segmentation using deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 3523–3542 (2021).
52. Laine, S. & Aila, T. Temporal ensembling for semi-supervised learning. In: *International Conference on Learning Representations* (2017).
53. Twinanda, A. P. et al. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* **36**, 86–97 (2016).
54. Nwoye, C. I. et al. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Med. Image Anal.* **78**, 102433 (2021).
55. Papandreou, G., Chen, L.-C., Murphy, K. P. & Yuille, A. L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision*, 1742–1750 (2015).
56. Olsson, V., Tranheden, W., Pinto, J. & Svensson, L. Classmix: Segmentation-based data augmentation for semi-supervised learning. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1378 (2021).
57. Lou, A., Tawfik, K., Yao, X., Liu, Z. & Noble, J. Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation. In: *IEEE Transactions on Medical Imaging* (2023).
58. Zou, Y. *et al.* PseudoSeg: Designing pseudo labels for semantic segmentation. In: *International Conference on Learning Representations* (2021).

Acknowledgements

Meng Wei is supported by the UKRI EPSRC CDT in Smart Medical Imaging [EP/S022104/1]. This work was supported by core funding from the Wellcome Trust / EPSRC [WT203148/Z/16/Z; NS/A000049/1]. Tom Vercauteren is supported by a Medtronic / RAEng Research Chair [RCSRF1819\7\34]. For the purpose of open access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. Tom Vercauteren is a co-founder and shareholder of Hypervision Surgical.

Author contributions

MW: Conceptualization, Methodology, Software, Data curation, Writing - Original draft preparation, Visualization, Investigation. CB: Software, Validation. LCGPH: Conceptualization, Writing - Review & Editing. RB: Software, Writing - Review & Editing. MS: Conceptualization, Methodology, Writing - Review & Editing, Supervision. TV: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

Additional information

Correspondence and requests for materials should be addressed to M.W. or M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025