Medical Image Analysis 97 (2024) 103246

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Reducing annotating load: Active learning with synthetic images in surgical instrument segmentation

Haonan Peng [a,*], Shan Lin [b], Daniel King [a], Yun-Hsuan Su [c], Waleed M. Abuzeid [a], Randall A. Bly [a], Kris S. Moe [a], Blake Hannaford [a]

[a] University of Washington, 185 E Stevens Way NE AE100R, Seattle, WA 98195, USA
[b] University of California San Diego, 9500 Gilman Dr, La Jolla, CA 92093, USA
[c] Mount Holyoke College, 50 College St, South Hadley, MA 01075, USA

ARTICLE INFO

MSC:
68T45
68T40
68U10

Keywords:
Robot instrument segmentation
Medical image synthesis
Active deep learning

ABSTRACT

Accurate instrument segmentation in the endoscopic vision of minimally invasive surgery is challenging due to complex instruments and environments. Deep learning techniques have shown competitive performance in recent years. However, deep learning usually requires a large amount of labeled data to achieve accurate prediction, which poses a significant workload. To alleviate this workload, we propose an active learning-based framework to generate synthetic images for efficient neural network training. In each active learning iteration, a small number of informative unlabeled images are first queried by active learning and manually labeled. Next, synthetic images are generated based on these selected images. The instruments and backgrounds are cropped out and randomly combined with blending and fusion near the boundary. The proposed method leverages the advantage of both active learning and synthetic images. The effectiveness of the proposed method is validated on two sinus surgery datasets and one intraabdominal surgery dataset. The results indicate a considerable performance improvement, especially when the size of the annotated dataset is small. All the code is open-sourced at: https://github.com/HaonanPeng/active_syn_generator.

1. Introduction

Minimally invasive surgery (MIS) has seen rapid development in recent years in applications such as intra-abdominal surgery and otolaryngology, and can improve surgical outcomes while reducing surgical morbidity (Sayari et al., 2019; Peters et al., 2018). In MIS, endoscopes are commonly used to provide visualization of the surgical site in real time. Segmentation of instruments is critical to the interpretation of endoscopic surgical images, and deep learning has been applied to this task (Maier-Hein et al., 2017; Shvets et al., 2018; Qin et al., 2020; Islam et al., 2019; Kalinin et al., 2020). In medical practice, labeled data is costly, and typically only trained experts can accurately annotate the images (Cheplygina et al., 2019; Yang et al., 2017). Consequently, small training sets are a frequent challenge to surgical instrument segmentation (Bodenstedt et al., 2021).

Recent efforts use synthetic data to alleviate the workload of annotating (Wang et al., 2021; Rajotte et al., 2021; Fujita et al., 2020). Synthetic images generated from simulation have accurate labels without manual work, but Su et al. (2021) suggest that the domain gap between synthetic and real datasets reduces the performance of a model trained by synthetic images. Domain adaptation techniques, such as generative adversarial networks (GANs) (Tobin et al., 2017; Kouw and Loog, 2019), can alleviate the domain gap and artifact of synthetic images. However, for endoscopic sinus surgery, because of reflections on metallic instruments, as well as blur and liquids on the tissue–instrument boundary, GAN-based methods may generate inaccurate synthetic images while the ground truth segmentation masks remain the same, which may confuse the training (Lin et al., 2020).

Besides GAN-based methods, the method of generating synthetic images by copying and pasting real object images onto real background images generalizes to many segmentation tasks (Ghiasi et al., 2021; Remez et al., 2018) and efficiently generates synthetic images with reduced concern for domain gap and inaccurate ground truth (Dwibedi et al., 2017), thereby enhancing the reliability and effectiveness of segmentation models. Furthermore, when background images are difficult to acquire, image inpainting methods such as patch-based image synthesis (Lee et al., 2016) can generate vivid backgrounds from cropped images.

When the resources for annotating real images are limited, the effectiveness of blending synthetic images through the copying and

* Corresponding author.
E-mail address: penghn@uw.edu (H. Peng).

https://doi.org/10.1016/j.media.2024.103246
Received 29 December 2022; Received in revised form 28 May 2024; Accepted 17 June 2024
Available online 22 June 2024
1361-8415/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

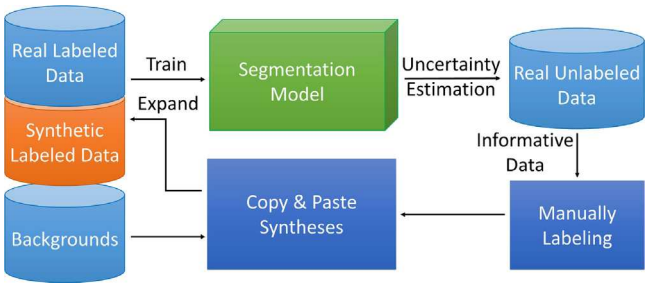

Fig. 1. Overview of the proposed method — iterative active learning with the generation of image syntheses. During the iterations, informative images are chosen from an unlabeled set by active learning. Synthetic images are generated from these selected images. Both real labeled and synthetic labeled images are used to train the segmentation model again to start a new iteration.

pasting method depends on the strategic selection of real labeled images. Active learning (AL) enhances this process by selecting the most uncertain or informative samples from an unlabeled dataset for manual annotation (Gorriz et al., 2017), thus efficiently reducing the reliance on extensive real labeled data (Budd et al., 2021; Angluin, 1988). When AL is combined with deep learning techniques, it facilitates faster convergence and achieves superior performance with less labeled data.

In this work, we develop a feasible method that combines AL with copy-and-paste image syntheses (Fig. 1), overcome the difficulty in the generation of synthetic images for endoscopic sinus surgeries (Su et al., 2021; Lin et al., 2020), and further improve the performance of instrument segmentation. We utilize active learning to choose informative unlabeled images to annotate, and a copy-and-paste method to generate synthetic images that have instruments and tissue backgrounds inherited from original real images, which can improve the utilization of the selected real images. We aim to test the hypothesis that the segmentation model trained with the synthetic images and a smaller number of selected real images has competitive performance compared to models trained on fully labeled real datasets. We also aim to measure the relative effectiveness of different types of synthetic images, external backgrounds, and realism on instrument segmentation accuracy. We also investigate the effects of fusion near the boundary of the instrument and multi-blending on the visible boundary artifact on the performance of segmentation near the boundary. Three open-source datasets are used in the experiments — the UW-Sinus-Surgery Cadaver Dataset, Live Dataset (Qin et al., 2020), and the EndoVis 2017 Dataset (Allan et al., 2019).

2. Related works

2.1. Generation of synthetic images

Using synthetic images is an intuitive approach to reducing the annotation workload. Dwibedi et al. (2017) proposed a similar 'cut' and 'paste' method to generate synthetic data for kitchen object detection. Their result showed that simply copying and pasting could result in artifacts such as aliasing of boundaries, which decreased learning performance. By improving the blending between sprite and background, their approach reached competitive performance combined with 10% real data. However, Ghiasi et al. (2021) showed that pasting without any blending had a similar performance to blending.

In further related studies, Remez et al. (2018) described object instance segmentation with weakly-supervised cut-and-paste adversarial learning: a discriminator was used to distinguish between real and synthetic images. GANs (Goodfellow et al., 2020) have been also implemented to generate synthetic medical images (Singh and Raza, 2021; Yoo et al., 2020) based on a similar discriminator. Recent implementations include image-to-image translation from simulated images to real images for cataract surgery (Luengo et al., 2018) and laparoscopic surgery (Colleoni and Stoyanov, 2021), as well as from cadaver images to live images (Lin et al., 2020) for sinus surgery.

2.2. Active learning

Active learning is dedicated to selecting and labeling the most informative training images that can reach near-optimal performance with the fewest annotations (human effort) (Tajbakhsh et al., 2020; Kim et al., 2020). Typically, in active learning, unlabeled images are selected by criteria such as maximum entropy and least confidence (Holub et al., 2008; Roels and Saeys, 2019; Schein and Ungar, 2007). In some cases, however, these criteria do not outperform random selection (Yang et al., 2017; Belharbi et al., 2021). Thus, more advanced criteria such as Bayesian active learning by disagreement (BALD) are proposed (Houlsby et al., 2011). The BALD criterion combines a high overall uncertainty with a term that increases the weight of disagreement among the population. Gal et al. (2017) presented a study on semantic segmentation of prostate medical images with active learning, and the BALD criterion outperformed maximum entropy, especially when the budget for annotation was small. Tran et al. (2019) proposed a Bayesian generative active deep learning, which combined active learning and GAN data augmentation. The evaluation of image classification tasks suggested that the combined method outperformed each single method. Bodenstedt et al. (2019) developed an active learning approach based on Deep Bayesian Networks for instrument presence detection and surgical phase segmentation. The experiments suggested that with the same amount of training data, active learning outperformed random selection in training the surgical workflow analysis model.

2.3. Unsupervised & semi-supervised learning

With plenty of unlabeled data but expensive annotation, unsupervised and semi-supervised learning are also preferred in medical imaging (Barragán-Montero et al., 2021; Raza and Singh, 2021). Liu et al. (2020) proposed an unsupervised learning method for surgical instrument segmentation that used generated anchors as pseudo labels with ambiguity resolved by temporal coherence. Instead of using pixel-level annotations, Fuentes-Hurtado et al. (2019) developed an approach utilizing weak annotations provided as stripes over the different objects in the image. Recent efforts also include utilizing the local center of mass (Aganj et al., 2018), 2D points of interest (Lejeune et al., 2018), and existing annotations from other datasets (Sestini et al., 2023). With minimized usage of annotation, unsupervised and semi-supervised learning can achieve comparable performance to fully supervised learning. Momentum Contrast (MoCo) (He et al., 2020) for unsupervised visual representation learning also suggested close performance to supervised representation learning on medical imaging datasets (Ramesh et al., 2023; Hirsch et al., 2023). Besides the current frame, external information such as robot pose (Qin et al., 2019; Sestini et al., 2021) and coherent frames (Funke et al., 2018; Lin et al., 2021) can also be utilized to improve the performance of segmentation in supervised and self-supervised learning models.

Among approaches for the generation of accurate, labeled synthetic images in endoscopic surgery, we leverage active learning to minimize the workload of manual segmentation, then copy surgical instruments and paste them into a surgical background image at random positions and orientations. The background image can be a video frame where a human confirms that no instrument is present, or it can be created from a segmented image by infill from adjacent backgrounds at the instrument location. This image synthesis method is relatively easy to implement and adjust. In comparison, while unsupervised learning eliminates the need for annotating training data, it may incur effort and cost to acquire pseudo labels or labeled data from other datasets and simulations. In contrast, our proposed method, requiring a small amount of labeled data, demands no additional human effort. By adjusting several parameters, this method can be adapted to different datasets (Fig. 5).

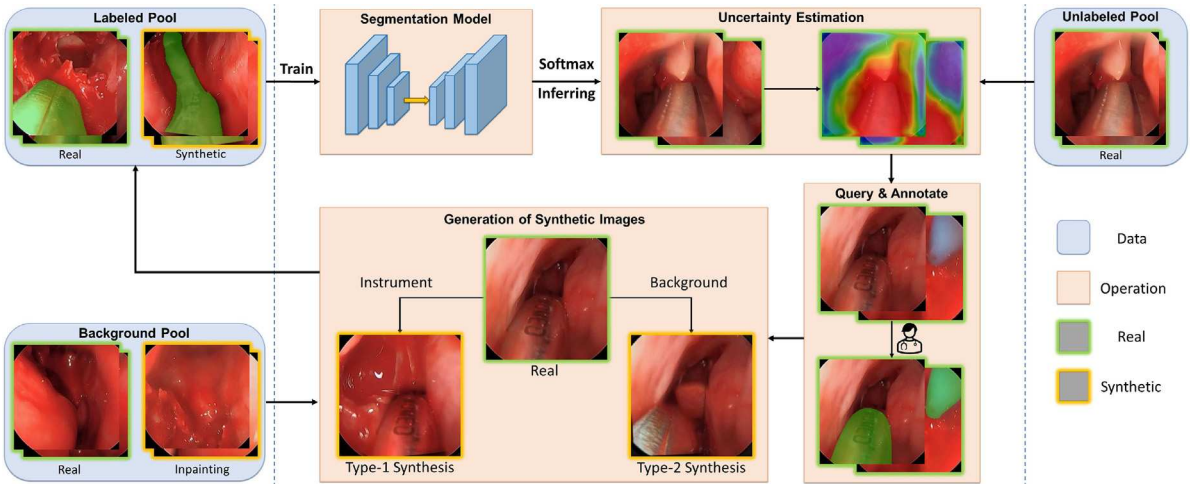

Fig. 2. Workflow of the system. The segmentation model is first trained by labeled images. Uncertainty estimation is performed on the unlabeled real images to select the most informative ones, which are then manually labeled. Next, synthetic images are generated from the labeled images. There are 2 types of synthetic images. Type-1 syntheses have the same instrument as the original real image while Type-2 syntheses have the same background.

3. Methods

3.1. System workflow

Fig. 2 shows the workflow of the entire system. During training, the proposed system uses active learning to choose the most informative samples from the unlabeled pool of the database and asks for annotation (Section 3.4). Then, synthetic images are generated and added to the labeled pool with the real images to 'make the best use of' the selected real images (Section 3.2). The goal is to train a segmentation model on fewer hand-labeled real images while maintaining competitive performance.

The dataset (Section 4.1) consists of a labeled pool and an unlabeled pool of images. Because endoscopes are widely used in robot-assisted surgeries, it is not difficult to obtain unlabeled videos and images. Some of the datasets contain background images in which surgical tools are not present, which can form a background pool.

Initially, some real images are randomly chosen and moved from the unlabeled pool to the labeled pool by human annotation. Synthetic images are first generated using the images in the labeled pool and are then added back to the labeled pool. Next, a segmentation model is trained using the labeled pool. Then, uncertainty estimation is applied to the unlabeled pool, and the most informative images are queried by the BALD active learning criterion, asking for annotation. Synthetic images are generated based on the newly labeled images. For each real image, there are two types of synthetic images. Type-1 synthetic images have the same surgical tool as the original real image, and the background is randomly selected from the background pool. Type-2 synthetic images have the same background as the original real image (background inpainting is applied to the original real image to remove the original tool), and the surgical tool is randomly selected from the labeled pool. If the dataset has no or few background images, background inpainting of instrument pixels in the selected labeled real images is performed and the generated backgrounds are added to the background pool (Section 3.3).

After the generation, the newly labeled real images and the synthetic images are added to the labeled pool and then the segmentation model is trained again to start a new iteration. This is repeated until the labeling budget or the desired performance is reached. Budget in this paper is defined as the fraction of real training images which are manually annotated, compared to the total number of real training images.

3.2. Generation of synthetic images

The generation of each synthetic image begins with one selected real labeled image. To generate Type-1 synthetic images, the selected image is used as the instrument image, and a background image is randomly chosen from the background pool. To generate Type-2 synthetic images, the selected image is used as the background image with inpainting applied to remove the instrument, and an instrument image is randomly selected from the labeled pool. The tool from the instrument image is copied and pasted on the background image, with resizing, movement, and fusion. Fig. 3 shows the workflow of the generation of synthetic images.

The procedure starts with x^i - a labeled real image that includes an instrument, and x^b - a pure background. The instrument image x^i also has a mask y^i, a binary matrix with the same size as x^i in which the instrument pixels are 1 while other pixels are 0. Resizing, movement, and rotation are first applied to the instrument and the mask:

$$x_r^i = R(x^i, c, w, h, \theta) \tag{1}$$

$$y_r^i = R(y^i, c, w, h, \theta) \tag{2}$$

where $R(\cdot)$ is the operator, c, w, h, and θ are the factors of resizing, movement in width and height, and rotation angle, respectively. These operations are applied sequentially. A binary dilation (Dougherty, 1992) is applied on the new mask y_r^i so that in the dilated mask y_d^i, the region of the instrument is larger than the true mask of the instrument y_r^i.

$$y_d^i = \bigcup_{b \in B} y_{rb}^i \tag{3}$$

where B is a $d \times d$ matrix and d is the dilation kernel size, y_{rb}^i is the translation of y_r^i by b. After dilation, the fusion mask y^f is generated by applying average blur or Gaussian blur (Young and Van Vliet, 1995) to the mask y_d^i:

$$y^f = B_a(y_d^i, k) \tag{4}$$

$$y^f = B_G(y_d^i, k, \sigma) \tag{5}$$

where $B_a(\cdot)$ and $B_G(\cdot)$ are the operators of average blur and Gaussian blur, respectively. k is the kernel size and σ is the standard deviation. According to Dwibedi et al. (2017), the artifacts from the copy-and-paste operation may result in decreased performance if the model is trained on synthetic images. Multi-blending, using the same synthetic

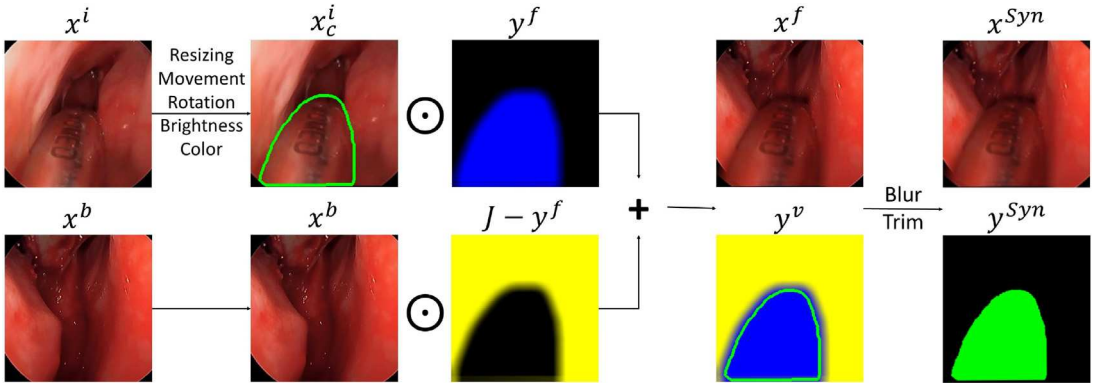

Fig. 3. Generation of a synthetic image. Note that y^v is only for visualization, where the solid green line indicates the outline of the instrument, the yellow area is from the background image x^b and the blue area is from the instrument image x^i_c. The transition area can be found around the boundary of the instrument on the synthetic image.

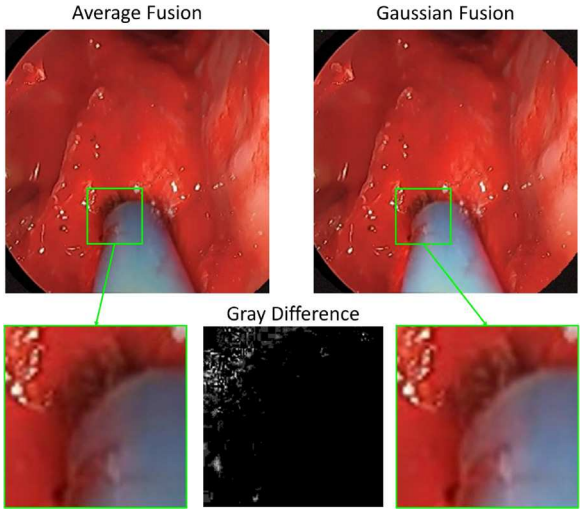

Fig. 4. Multi-blending: images on the left and right are similar because they have the same size and position of the instrument and background, however, the instruments are blended on the background by different blending methods and parameters (average fusion on the left and Gaussian fusion on the right). According to Dwibedi et al. (2017), when training deep learning models, using multi-blended synthetic images can prevent the negative effect of blending artifacts.

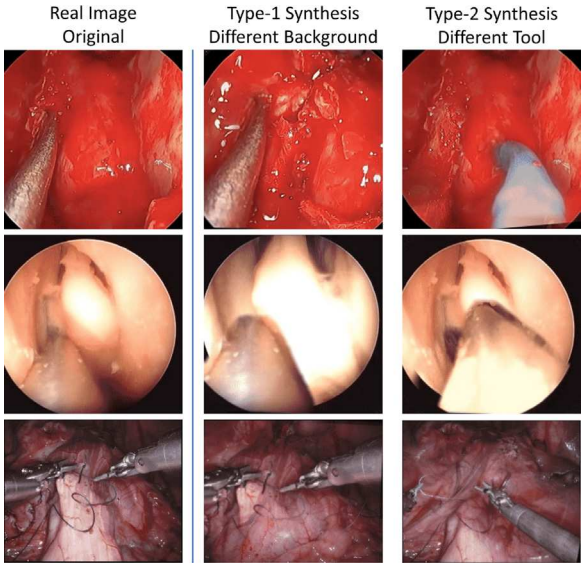

Fig. 5. Original real images (left), Type-1 synthetic images (center) and Type-2 synthetic images (right). Compared with the real images, Type-1 has different backgrounds and Type-2 has different instruments.

images but different blending methods in the training set, can prevent the model from learning blending artifacts and improve the performance on real images (Fig. 4). Color and brightness adjustment is applied to narrow the gap between the color style of the instrument and the background, for each channel of the adjusted image:

$$x^i_{c,chn} = \beta \frac{\sum x^b}{\sum x^i_r} \left(\alpha \frac{\sum x^b_{chn}}{\sum x^i_{r,chn}} x^i_{r,chn} + (1-\alpha) x^i_{r,chn} \right) \tag{6}$$

where α is the factor of color adjustment, β is the factor of brightness adjustment, x^b_{chn} and $x^i_{r,chn}$ are the same channel of the background image and the instrument image, respectively. After the adjustment of color and brightness, the instrument is blended onto the background by:

$$x^f = y^f \odot x^i_c + (\mathcal{J} - y^f) \odot x^b \tag{7}$$

where $\odot$ indicates element-wise multiplication, $\mathcal{J}$ is a matrix of ones with the same size as y^f. A weak Gaussian blur is applied and the border is trimmed to restore the outline, and thus finalize the generation of the synthetic image x^{syn} and the corresponding mask y^{syn}. Fig. 5 shows examples of Type-1 and Type-2 synthetic images.

3.3. Inpainting of backgrounds

As introduced in 3.2, a synthetic image is generated from a labeled instrument image and a background image. However, it is not always feasible to find background images in every dataset. Thus, for those datasets without background images, image inpainting is performed to generate backgrounds from labeled instrument images.

Fig. 6 shows the procedure of background inpainting. It is similar to the generation of synthetic images. The difference is that for the inpainting of backgrounds, an area of background is blended over the instrument pixels, instead of blending an instrument over a background. There are 2 types of inpainting, self-inpainting and external inpainting. First, self-inpainting can be performed by self-flipping or rotation of the original image (8), if the flipped or rotated mask does not overlap with the original mask (9):

$$x^b_i = y_f \odot x^i_r + (\mathcal{J} - y^f) \odot x^i \tag{8}$$

$$y^f \cap y^f_r = 0 \tag{9}$$

where x^i is the image including instrument, x^b_i is the inpainting background with instrument removed, y^f is the fusion mask generated by the method mentioned in 3.2, x^i_r and y^f_r are the flipped or rotated image

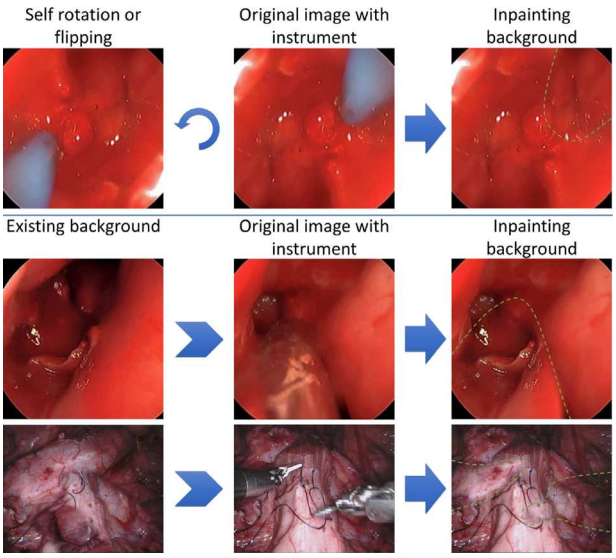

Fig. 6. Generation of inpainting backgrounds (right). The instruments on the original images (middle) are removed by self-rotation/flipping or existing backgrounds (left).

x^i and mask y^f. Flipping can be applied vertically or horizontally, and rotation can be applied for degrees of 90, 180, and 270.

However, sometimes the masks y^f and y_r^f always overlap regardless of flipping and rotation. In this case, external inpainting can be performed by randomly selecting another background from the background pool as a source for the pixels covering the instrument:

$$x_i^b = y^f \odot x^{b2} + (\mathcal{J} - y^f) \odot x^i \tag{10}$$

where x^{b2} is the background (original or inpainting) from the background pool and the other variables are the same as self-inpainting. To prevent the rare situation that self-inpainting is not applicable for all active selected real images in the first active learning iteration, 1 external background should be added to the background pool at the beginning of active learning. As active learning and generation of synthetic images proceeds, inpainting backgrounds will be added to the background pool.

3.4. Active learning

Active learning effectively reduces the annotation burden while ensuring that the model retains competitive performance with limited labeled data. The proposed system uses a pool-based AL method (Fig. 2). The iteration of the active learning is introduced in Section 3.1 and BALD is used as the criterion to query unlabeled images.

The BALD criterion chooses the images which are expected to maximize the mutual information between predictions and model posterior:

$$\mathbb{I}[y, \omega | x, \mathcal{D}_{train}] = \mathbb{H}[y|x, \mathcal{D}_{train}] - \mathbb{E}_{P(\omega|\mathcal{D}_{train})}[\mathbb{H}[y|x, \omega]] \tag{11}$$

where x is the input image and y is the output label, $\mathbb{H}[y|x, \mathcal{D}_{train}]$ and $\mathbb{H}[y|x, \omega]$ are the entropy (Shannon, 1948) of the prediction $P(\omega|\mathcal{D}_{train})$ and distribution $P(y|x, \omega)$, respectively. $\mathcal{D}_{train}$ is the labeled training data, and ω are model weights. The first term seeks the images which have high average entropy in the sampled models. The second term imposes a penalty such that the images on which the models disagree are kept while overall unconfident images are dropped.

To perform BALD, Monte-Carlo (MC) dropout is performed during training and inference. The implementation of the BALD criterion on image semantic segmentation is:

$$\mathbb{I}[y, \omega | x, \mathcal{D}_{train}] \approx -\sum_c \left(\frac{1}{T}\sum_t p_c^t\right) \log\left(\frac{1}{T}\sum_t p_c^t\right) + \frac{1}{T}\sum_{t,c} p_c^t \log p_c^t \tag{12}$$

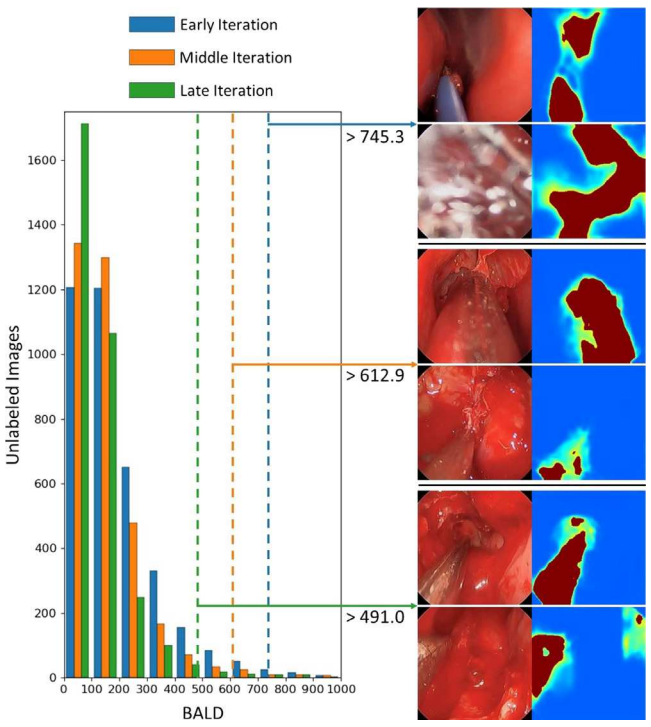

Fig. 7. The distribution of BALD acquisition values on unlabeled images (left) and examples of queried images with the softmax probability masks (right) in early, middle, and late AL iterations. 395 real annotated images were used in the training and generation of synthetic images. The segmentation model was retrained for 70 epochs after each active learning iteration.

where c is the number of classes, T is the number of committee members (models trained and inferred with MC dropout), and p_c^t is the softmax probability of the pixel. Examples of BALD distribution and queried unlabeled images are shown in Fig. 7.

4. Experiments and results

4.1. Datasets

All the images in the 3 datasets below were manually labeled by experts (attending surgeons and surgical residents). However, ground truth masks were hidden by default and were provided only when the images were in the test set or were marked as 'labeled'.

UW-Sinus-Surgery-C/L Dataset (Qin et al., 2020) contains two parts: the live dataset (Sinus-Live) and the cadaver dataset (Sinus-Cadaver). For the Sinus-Live dataset, 3955 labeled images from the first two videos were used as the training set. And the third video contains 701 labeled images that were used as the test set. 696 background images were manually selected from the first two videos and provided to the system when external backgrounds are required. Manually choosing backgrounds was not as costly as annotating the segmentation of images. Three non-medical laypersons were asked to select 700 backgrounds out of 20 000 images, which took 25 min, 17 min, and 31 min, respectively. All the images were resized and center-cropped to 240 × 240.

The images in the test set and training set were from different videos of different surgical procedures. Because there were very few real images used in some experiments, non-informative images, such as pure black or white images caused by over-exposure or blocking, were manually removed from the training set. However, no image was removed from the test set to ensure that the performance was fairly evaluated.

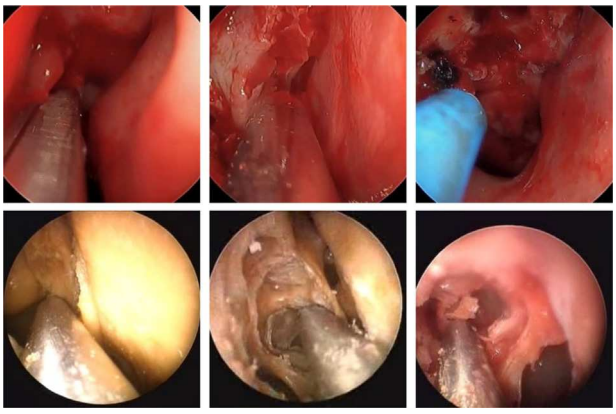

Fig. 8. Example images from Sinus-Live (top) and Sinus-Cadaver (bottom) datasets.

The Sinus-Cadaver dataset was built similarly, collected from 10 surgery videos on 5 cadaver specimens. The training set, test set, and background set have 2908, 1437, and 597 images, respectively. Due to the different conditions of each cadaver specimen, the overall appearance of the images can be different. However, none of the recorded videos show considerable similarity to real surgeries. Humans have no difficulty visually distinguishing cadaver videos and live videos. Examples from Sinus-Live and Sinus-Cadaver can be found in Fig. 8.

EndoVis 2017 Robotic Instrument Segmentation Dataset is from one of the sub-challenges of MICCAI 2017 (Allan et al., 2019). The images were derived from 10 sequences of abdominal porcine procedures recorded using da Vinci Xi robotic endoscopic surgery systems. The instruments used include Large Needle Driver, Prograsp Forceps, Monopolar Curved Scissors, Cadiere Forceps, Bipolar Forceps, Vessel Sealer, and an ultrasound probe. The selected frames were labeled by a segmentation team at Intuitive Surgical. Although the videos were recorded by stereo cameras, only left-eye images were labeled. We used 900 images (225 for each video) with labels, from videos 1–4, as the training set. 900 images and labels from videos 5–8 were used as the test set. Due to the long training time of the active learning, the images were resized to 427 × 240 to reduce computational time.

4.2. Parameters of generation of synthetic images

The workflow of the generation of image syntheses is shown in Fig. 3. The parameters of the generation of synthetic images were empirically chosen and applied to all the experiments unless stated otherwise. (Type-1, Type-2) syntheses per selected real image were set to (2, 0) without multi-blending for the Sinus-Live and Sinus-Cadaver datasets, and (0, 1) with multi-blending for the EndoVis 2017 dataset. External backgrounds were only provided to the Sinus-Live and Sinus-Cadaver datasets, while background inpainting was only applied to the EndoVis 2017 dataset. The factor of tool resizing c was set to range [0.9, 1.2] for all datasets. The movement width w and height h were set to up to 24 pixels, with rotation θ up to 30 degrees. When blending instruments on backgrounds, the dilation kernel d was set to 15 pixels for all datasets, while the fusion blur kernel k was set to range [10, 15] pixels for Sinus-Live and EndoVis 2017, and range [5, 10] for Sinus-Cadaver. The factor of color adjustment α was set to range [0.4, 1.0]. A larger α results in a stronger adjustment and 0 means no adjustment. The range of brightness adjustment β was [0.9, 1.3]. The larger β, the brighter the adjusted image is, and 1.0 means no adjustment.

4.3. Training details

The segmentation model used in this paper has the same structure as Qin et al. (2020), unless stated otherwise. This is a modified DeepLabv3+ (Chen et al., 2018) encoder–decoder model with MobileNet (Howard et al., 2017) as the feature extractor. To fit in the active learning iterations, the learning rate was increased, and the training iterations were decreased significantly to accelerate the training, though the accuracy was slightly compromised. Adam (Kingma and Ba, 2014) was used as the optimizer, and the exponential decay rates of the 1st and 2nd order moment estimates were 0.9 and 0.999, respectively. The batch size was set as 16. Because we used different budgets of the hand-labeled training set in the experiments, for all the experiments in this paper, the training iterations were 20 basic epochs plus a compensation of 5/budget(percentage) epochs to ensure convergence when a small budget is used. For example, epochs were 25 (20 + 5) for 100% real images budget, and 30 (20 + 10) for 50% budget. The segmentation model was retrained for the epochs after each active learning iteration. The initial learning rate was set to 0.001 and exponential decay strategy was applied. The backbone lightweight MobileNet was pretrained on ImageNet (Russakovsky et al., 2015). Image augmentation was also applied to the training data for better generalization ability, which includes hue, brightness, saturation, contrast, flipping, rotation, zooming, and zero-padding.

In the beginning, all the images in the dataset were in the unlabeled pool and the labels were hidden. For example, if the real-image budget was 394 images (10% of the Live dataset), then 197 (half) real images were randomly chosen first, and the rest 197 images were chosen by the BALD criterion in 3 iterations. Once chosen from the unlabeled pool, the real images were moved to the labeled pool and their labels were revealed. Synthetic images were generated with labels naturally.

4.4. Evaluation metrics

Two main evaluation metrics were used in this paper, Dice similarity coefficient (DSC) and intersection over union (IoU) (Taha and Hanbury, 2015), which are defined as

$$DSC = \frac{2|S \cap G|}{|S| + |G|}, IoU = \frac{|S \cap G|}{|S \cup G|}$$

where S is the foreground pixels of prediction, G is the corresponding ground truth, and $|*|$ is the counting operation. To study the effect of blending and fusion on the performance of segmentation near the boundary, IoU near boundary (IoU_{NB}) was used as an additional metric,

$$IoU_{NB} = \frac{|S \cap G| \cap B}{|S \cup G| \cap B}$$

where B denotes the near-boundary binary mask with width of 20 pixels band region near the instruments' boundary. The mean values of these three metrics are calculated over each test, denoted as mDSC, mIOU, and $mIoU_{NB}$.

4.5. Usage of real images

The performance of the proposed method was evaluated when different annotation budgets were used. Budgets were set as proportions of the total real images in the training set as introduced in Section 4.1, from 1% to 100%. For each budget, 4 tests were performed - (1) randomly chosen training images without synthetic images, (2) BALD implemented - half chosen by BALD and the other half chosen randomly, without synthetic images, (3) randomly chosen real images with generated synthetic images, (4) BALD implemented with generated synthetic images. We also compared the proposed method with baseline results, which were obtained when the training budget was 100% with no synthetic images.

The evaluation results of segmentation performance are shown in Table 1 and method ablation on more budgets is shown in Fig. 9. Each entry is the average of 5 repetitive tests with different global random seeds to reduce the influence of randomness, which applies to all the experiments in this paper. From Fig. 9, it can be seen that by generating synthetic images with active learning, the performance of segmentation

Table 1
Segmentation performance with different budgets.

Budget	Proposed		Performance (%)								
	BALD	Syn	Sinus-Live			Sinus-Cadaver			EndoVis 2017		
			mDSC	mIoU	$mIoU_{NB}$	mDSC	mIoU	$mIoU_{NB}$	mDSC	mIoU	$mIoU_{NB}$
5%	×	×	69.31	61.83	54.00	68.97	60.57	53.04	76.89	64.50	62.44
	✓	✓	**72.41**	**65.86**	**56.96**	**75.96**	**69.12**	**59.09**	**81.56**	**71.09**	**71.28**
20%	×	×	74.05	66.59	58.19	73.45	65.70	55.58	80.90	70.10	71.16
	✓	✓	**76.78**	**70.43**	**61.82**	**76.58**	**70.27**	**60.73**	**83.05**	**73.25**	**73.20**
50%	×	×	77.93	71.26	62.83	75.36	68.31	58.18	81.70	71.04	71.40
	✓	✓	**80.66**	**74.87**	**66.30**	**78.64**	**72.49**	**62.49**	**83.01**	**73.32**	**73.45**
100%	×	×	81.35	75.14	66.34	79.42	72.50	62.61	82.50	72.52	**74.28**
	×	✓	**83.64**	**78.18**	**70.86**	**80.28**	**73.83**	**62.84**	**84.04**	**74.66**	74.21

(i) The bold font indicates the best performance in each budget. (ii) BALD active learning is not applicable when the budget is 100%.

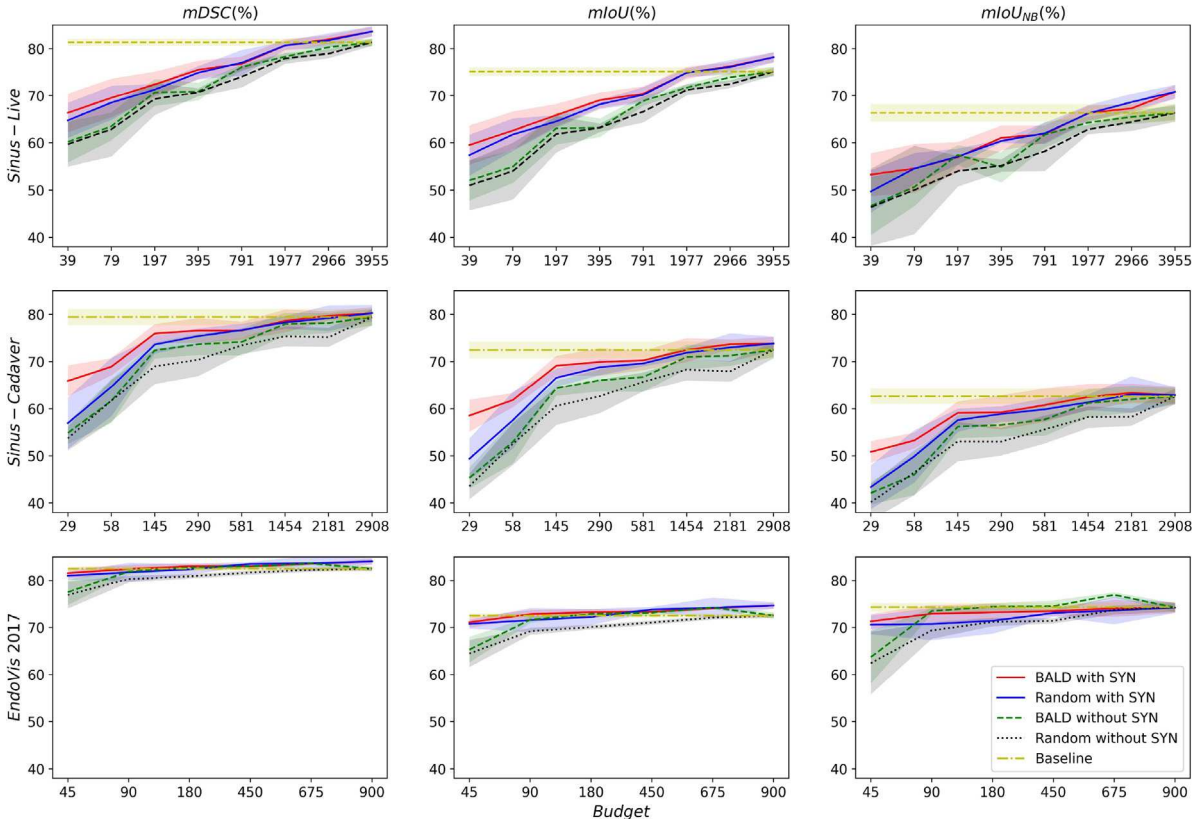

Fig. 9. Evaluation of the model trained by different budgets of real images, with and without active learning (BALD) and synthetic images (SYN). The solid lines indicate the average over 5 repetitive tests and the shades indicate the standard deviation. The horizontal axes are nonlinear for better visualization. The training on the EndoVis 2017 dataset did not converge on the 1% or 2% budget because of too few images. Thus, the evaluation of this dataset started at 5%. The baseline result is obtained by training the model on 100% real images (no synthetic images), regardless of budget.

was significantly improved when the budget was small. Compared with randomly chosen real images and no synthetic images, the average improvement of the proposed method in mDSC with less than 10% budgets was 5.31%, 8.15% and 3.41% in Sinus-Live, Sinus-Cadaver and EndoVis 2017 datasets, respectively. Specifically, the improvement on the Sinus-Live and Sinus-Cadaver dataset with only 1% budget was 6.67% and 12.20%. Overall, as the budget increased, the improvement became smaller. However, a considerable improvement could still be achieved when the budget was 100% of the training set (BALD was not applicable). The improvement in EndoVis 2017 dataset was not as significant, but the proposed method began to outperform the baseline result at only 10% usage of the training set.

Similar trends can be seen in mIoU and mIoU near the boundary. The average improvement of the proposed method with small budgets (less than 10%) was significant in mIoU — 7.11%, 10.02%, and 4.93% in the three datasets, as well as $mIoU_{NB}$ — 5.22%, 7.42% and 6.39%. Considerable improvements in mIoU could still be seen when the budget is 100% of the training set — 3.04%, 1.33% and 2.14%. However, for $mIoU_{NB}$, only the result of the Sinus-Live dataset showed improvement (4.52%), while the results of the other two datasets were similar to the baseline results.

4.6. Number and type of synthetic images and multi-blending

As introduced in Section 3.2, there are 2 types of synthetic images. For each chosen and labeled real image, Type-1 synthetic images have the same instrument and Type-2 synthetic images have the same background as the real image. Multi-blending is also reported in Dwibedi et al. (2017) to avoid decreased performance caused by the artifact near the boundary in synthetic images. Thus, this experiment was

Table 2
Segmentation performance with different types of synthetic images and multi-blending.

Group	Syn per real image			Performance (%)								
	Type-1	Type-2	M-blend	Sinus-Live			Sinus-Cadaver			EndoVis		
				mDSC	mIoU	$mIoU_{NB}$	mDSC	mIoU	$mIoU_{NB}$	mDSC	mIoU	$mIoU_{NB}$
	No synthetic image			71.70	64.19	55.47	69.35	61.81	52.76	79.55	68.06	67.88
	1	1	1	74.74	68.24	59.34	73.32	66.89	56.15	81.94	71.77	71.76
	0.5	0.5	2	74.64	68.09	60.54	74.30	67.80	58.31	81.47	71.19	71.14
1	2	0	1	**76.97**	**70.47**	61.65	**76.59**	**69.93**	59.22	81.37	71.14	72.62
	1	0	2	74.75	68.50	60.16	75.29	68.45	**59.24**	81.34	71.35	**74.24**
	0	2	1	75.07	68.67	60.63	71.90	65.77	55.69	82.25	72.03	71.60
	0	1	2	76.80	70.38	**63.23**	72.85	66.44	56.34	**82.43**	**72.46**	73.31
2	4	4	1	**78.03**	**72.23**	**62.96**	70.86	65.33	54.84	80.90	**70.48**	68.50
	2	2	2	77.57	71.89	62.91	**76.09**	**70.32**	**59.67**	**80.96**	70.39	**68.57**
3	6	6	1	77.42	71.73	61.95	69.90	64.46	54.21	80.43	69.77	66.96
	3	3	2	**77.91**	**72.15**	**62.79**	**74.90**	**69.45**	**58.91**	**80.95**	**70.29**	**67.61**

The bold font indicates the best performance in each group. Under Syn per Real Image, the parameters of Type-1 and Type-2 indicate that for each labeled real image selected by the active learning mechanism, how many Type-1 and Type-2 synthetic images were generated, respectively. A value of 0.5 means that there is a 50% chance to generate a synthetic image. M-blend value of 1 means that each synthetic image is single and multi-blending is not applied. And M-blend value of 2 means that for each synthetic image blended by average fusion (4), there is another similar synthetic image blended by Gaussian fusion (5).

performed to study the effectiveness of the 2 types of synthetic images and multi-blending. To better compare the results, the experiments were separated into 3 groups. Each group tested the same number of generated images per real image. For example, in Table 2, tests in Group 2 featured 8 synthetic images for each queried real image — $(2[Type-1]+2[Type-2])\times 2[Multi-blending]$. Consequently, tests in each group had the same number of training iterations to ensure that the model was trained for the same fixed number of steps. To ensure convergence, instead of keeping training iterations, the training epochs of Group 2 and 3 were the same as Group 1 so that the training iterations of Group 2 and 3 were 4 and 6 times larger compared to Group 1, respectively. The annotation budget was fixed at 10% of the training set.

The results are shown in Table 2. In Group 1, for the 2 sinus surgery datasets, the best performance on mDSC and mIoU was achieved by 2 Type-1 synthetic images, while the best performance on $mIoU_{NB}$ was achieved by 1 Type-2 image with multi-blending on the Sinus-Live dataset, and by 1 Type-1 image with multi-blending on Sinus-Cadaver dataset. For the EndoVis 2017 dataset, 1 Type-2 image with multi-blending gave the best result on mDSC and mIoU, and 1 Type-1 image with multi-blending gave the best result on $mIoU_{NB}$. In Group 2, although the training steps were increased significantly compared to group 1, a decrease in performance could be seen in the Sinus-Cadaver and EndoVis 2017 datasets. Within the group, there was little difference in results with and without multi-blending on the Sinus-Live and EndoVis 2017 datasets. However, multi-blending increased the performance significantly in the Sinus-Cadaver dataset. A similar trend was observed in Group 3. Although no considerable difference was seen on mDSC and mIoU in Sinus-Live and EndoVis 2017 dataset by applying multi-blending, an improvement was observed on $mIoU_{NB}$.

4.7. External backgrounds

For the proposed method, backgrounds can be generated by inpainting from labeled instrument images or be provided from external backgrounds. This experiment studied whether providing external backgrounds can help with the segmentation result. The tests were separated into 4 groups according to the budget of real images. In each group, there are 4 sub-tests. One test only used real images to train the segmentation model without synthetic images. The remaining 3 sub-tests were all supplemented with synthetic images (BALD implemented). The only difference was how the backgrounds were provided. Because no external backgrounds (frames without instruments) could be found in the EndoVis 2017 dataset, all synthetic images for this dataset had to be generated from inpainted backgrounds. Thus, only the two sinus datasets were used in this experiment.

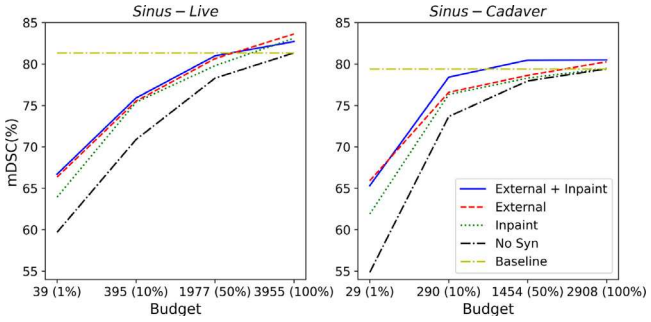

Fig. 10. Segmentation result (mDSC) with different backgrounds. The baseline result is obtained by training the model on 100% real images (no synthetic images).

The result in Fig. 10 (mDSC) shows that including external backgrounds improved the performance significantly when few real images were used (1% of the training set). For the Sinus-Live dataset, compared with no external backgrounds, the improvement was 2.77%, 2.89% and 1.93% in mDSC, mIOU and $mIoU_{NB}$, respectively. For the Sinus-Cadaver dataset, the improvement was 4.02% (mDSC), 4.86% (mIOU) and 1.74% ($mIoU_{NB}$). However, when a greater proportion of real images were used, the improvement was less considerable. For the Sinus-Live dataset, when 100% of the labeled real-image training set was used, the improvement was 0.52%(mDSC), 0.70%(mIOU) and 0.60%($mIoU_{NB}$). But for the Sinus-Cadaver dataset, although improvement could still be seen in mDSC and mIOU with external backgrounds, a decrease in performance was found in $mIoU_{NB}$(−0.99%).

4.8. Comparison with state-of-the-art

Fig. 11 presents a performance comparison of our proposed method with state-of-the-art techniques. On the Sinus-Live dataset, the proposed method outperformed GAN-based approaches (Su et al., 2021; Lin et al., 2020) with less annotated data used. The proposed method also achieved close performance with fully-supervised learning (Lin et al., 2021).

ResNet50 (He et al., 2016) and MobileNet (Howard et al., 2017) were used as the backbone model to evaluate the performance of the proposed method on the EndoVis 2017 dataset. With ResNet50 backbone, a small training set of 45 labeled images, and no other additional human efforts such as acquiring pseudo labels, our approach achieved competitive performance among the state-of-the-art unsupervised and semi-supervised learning approaches (Liu et al., 2020; Sestini et al., 2023; Ross et al., 2018), as well as the generation of

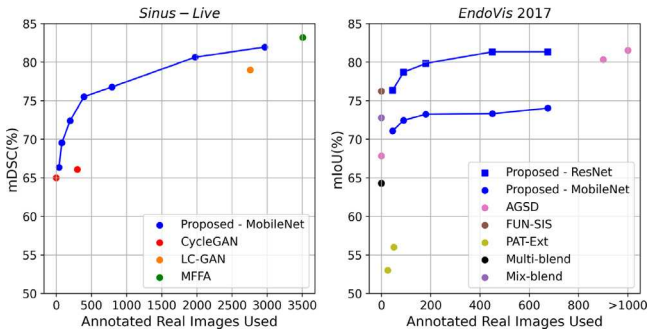

Fig. 11. Comparison of segmentation performance with state-of-the-art approaches. Sinus-Live dataset: CycleGAN (Su et al., 2021), LC-GAN (Lin et al., 2020), and fully supervised MFFA (Lin et al., 2021). EndoVis 2017 dataset: AGSD (Liu et al., 2020), FUN-SIS (Sestini et al., 2023), PAT-Ext (Ross et al., 2018), Multi-blending and Mix-blending (Dwibedi et al., 2017; Garcia-Peraza-Herrera et al., 2021). Evaluation metrics (mDSC or mIoU) are chosen based on the commonly used ones in the compared methods. Note that LC-GAN used annotated images from the cadaver dataset and FUN-SIS used annotated images from other existing datasets. Multi-blend and Mix-blend methods used 14 180 auto-labeled instrument foregrounds to generate synthetic images. For the Sinus-Cadaver dataset, to the best of the authors' knowledge, no unsupervised, semi-supervised, or GAN-based learning methods were reported at present.

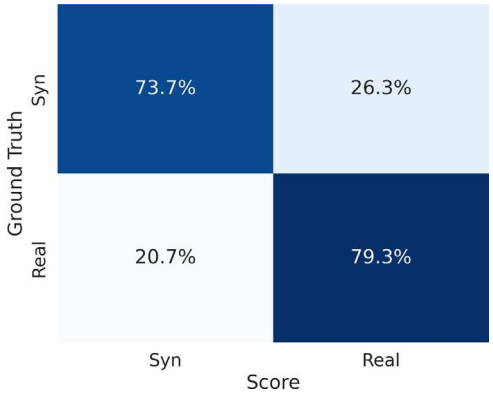

Fig. 12. Confusion matrix of the image review result from 3 expert surgeons. 200 synthetic and 100 real images were reviewed. For each image, experts were asked to choose whether the image was more likely to be 'real' or 'synthetic'.

synthetic images (Garcia-Peraza-Herrera et al., 2021). Compared to the accuracy of 50% supervised learning using 900 labeled images reported in AGSD (Liu et al., 2020), our method achieved better accuracy using 450 labeled images, and had a close performance as the fully supervised training in AGSD. When using MobileNet as the backbone model, our proposed method did not achieve the best accuracy of segmentation. However, compared to ResNet50, this lightweight model had faster training and inference speed and could be suitable if efficiency is a priority or computational resources are limited.

4.9. Realism of synthetic images

The proposed method generates synthetic images by placing instrument sprites according to multiple parameters including scaling, translation, rotation, blending, and color adjusting. To generate various synthetic images, these parameters are randomly chosen in relatively wide ranges. Some of the parameters (particularly edge blending) can affect the level of image realism apparent to humans.

To study how human-perceived realism impacts learning performance, the realism of the synthetic images needed to be evaluated. Three expert sinus surgeons participated in this experiment and performed a review of images. 100 real and 200 synthetic images were randomly chosen and generated from the UW Sinus-Live dataset. These images were shown to the participants and they chose whether the image appeared 'real' or 'synthetic'. Thus, the realism of a synthetic image could be evaluated by how many times it was considered 'real' by an expert (Fig. 12). Examples of the synthetic images that were unanimously rated real (3 of 3) and not real (0 of 3) are shown in Fig. 13.

Table 3
Segmentation performance with different syntheses realism.

Dataset	Images		Performance (%)		
	Real	Syn	mDSC	mIoU	$mIoU_{NB}$
Baseline	100	0	62.51	55.89	48.35
More Realistic	100	107	68.04	61.15	55.22
Less Realistic	100	107	**70.28**	**63.35**	**57.69**
All	100	200	68.60	62.30	57.08

In the review results, 10, 31, 66, 93 synthetic images received 3, 2, 1, and 0 realism scores, respectively (Fig. 14). To study the effectiveness of different levels of realism in synthetic images, 4 training datasets were formed: (1) Baseline: only the 100 real images; (2) More realistic: 100 real images and 107 synthetic images with at least 1 'real' rating; (3) Less realistic: 100 real images and 107 synthetic images (93 with 0 'real' rating, and 14 with 1 'real' rating, randomly chosen to compensate the size and thus the training iterations were the same); (4) All images: all 100 real image and 200 synthetic images.

The performance of models trained by these 4 training sets (Table 3) suggested that training with more realistic synthetic images or less realistic images had similar performance. The less realistic dataset slightly outperformed the more realistic dataset by 2.24%, 2.2%, and 2.47% in mDSC, mIoU, and $mIoU_{NB}$, respectively. Including all synthetic images achieved medium performance.

To further explore the impact of human-perceived realism on training effectiveness, feature extraction based on DINO (Caron et al., 2021) was performed on real images in the training and test set, as well as more and less realistic synthetic images. Augmentation was also applied, the same as the training of the segmentation models in this paper (details in Section 4.3). Then a principle component analysis (PCA) was followed to reduce the extracted features to 2D for easier visualization, presented in Fig. 15. In this PCA subspace, less realistic images displayed a marginally greater variety than their more realistic counterparts. In Fig. 15, the average distance of each image in the test set to the closest image in the more realistic training set was 3.217, while the distance of the less realistic training set was 3.211. In comparison, this average distance for the training set with only real images was 3.353. Thus, this experiment suggested that although the proposed copy-and-paste method could generate human-perceived unrealistic images, these images could still contribute to the training and enhance the variety of the training data.

5. Discussion

In this paper, we studied the use of image syntheses directed by active learning to improve the performance of surgical instrument segmentation. The results of the study on real image usage indicated that the proposed method improved the performance of segmentation significantly, especially when few real labeled images were used. With 10% real images budget combined with actively generated synthetic images, the performance of segmentation was comparable to using 50% of real images without synthetic images, cutting manual labeling effort by 80%. The performance of the proposed method was comparable with the baseline result (using 100% of the hand-labeled training set) when using only 50%, 75%, and 10% of hand-labeled data when evaluated on the Sinus-Live, Sinus-Cadaver and EndoVis 2017 datasets, respectively.

For the sinus surgery datasets, Type-1 synthetic images (the same instrument as the real image but with a different background) slightly outperformed Type-2 synthetic images (different instruments but with

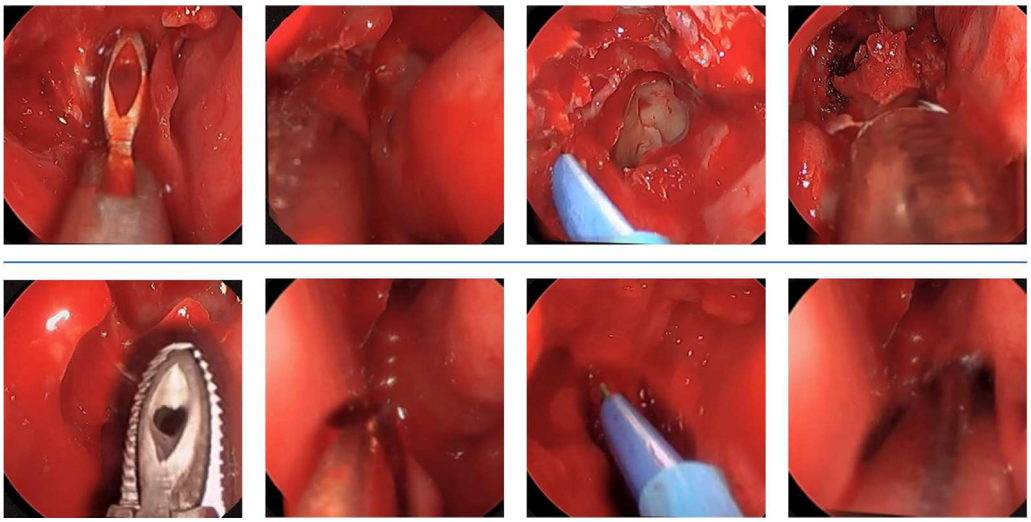

Fig. 13. Examples of the most realistic synthetic images that received 3 out of 3 realism ratings (top), and less realistic synthetic images received 0 out of 3 realism ratings (bottom).

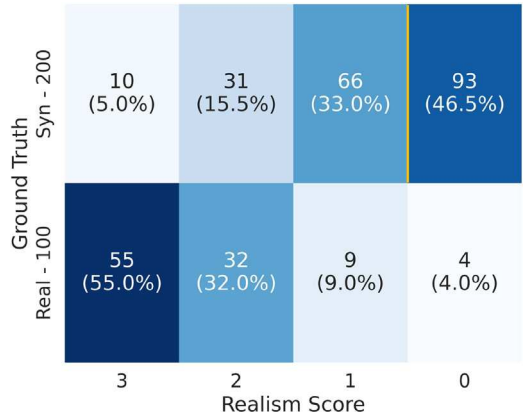

Fig. 14. Realism ratings from the image review. Synthetic images with 1 or more realism ratings are considered 'more realistic', and otherwise 'less realistic', as the yellow line indicates.

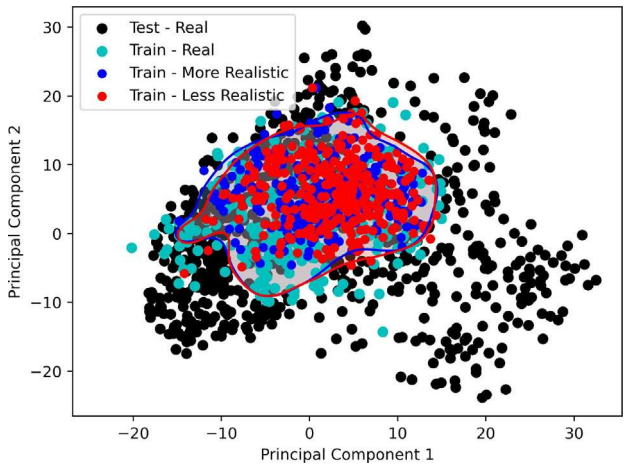

Fig. 15. Principle component analysis on extracted features of real images in the training and test set, more realistic synthetic images, and less realistic synthetic images, corresponding to Table 3. Despite that less realistic synthetic images expand the training set slightly better than more realistic synthetic images, the distributions of the two image sets are close. This is notable considering their significant differences in human-perceived realism, as depicted in Fig. 13.

the same background). Although multi-blending made no major improvement in overall performance, it improved the segmentation near the boundary. In addition, too many synthetic images were observed to decrease the performance (Group 3 of Table 2).

We varied the method of generating background images between inpainting the space occupied by instruments, and using original video frames that contained no instruments (external backgrounds). The use of external backgrounds improved the performance significantly when the annotating budget was extremely small, without adding much workload. Because manually selecting background-only frames from surgical videos requires no expertise and is not as time-consuming as labeling the instruments, especially when different parts of the instruments have different labels, adding external backgrounds is an efficient way to improve the segmentation with small numbers of labeled images.

6. Conclusion

Motivated by alleviating the experts' workload of annotating for instrument segmentation in endoscopic images, we propose the use of actively generated synthetic images to reduce the need for labeled real images while having comparable performance. When training segmentation models, the proposed method selects the most informative unlabeled images, then annotates these images and generates synthetic images based on the selected real images. Thus, a more diverse training set is formed by labeled real images and synthetic images, which results in considerable improvement in performance compared with using real images only, especially when the budget for annotating "new" images is small. The proposed method utilizes and combines active learning and generation of synthetic images to reduce the usage of real labeled images, and can be flexibly applied to different segmentation models and datasets, with different active learning criteria.

The current framework is limited to binary segmentation. In the future, we plan to generalize the proposed method to semantic segmentation among different instruments and parts. We also plan to further explore the relationship between realism to humans and effectiveness to artificial intelligence. There are many factors affecting the realism of synthetic images, and the influence of these factors on machine learning performance remains to be discovered.

CRediT authorship contribution statement

Haonan Peng: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Shan Lin:** Writing – review

& editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Daniel King:** Software, Methodology, Data curation. **Yun-Hsuan Su:** Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Waleed M. Abuzeid:** Writing – review & editing, Writing – original draft, Validation, Supervision, Formal analysis. **Randall A. Bly:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Conceptualization. **Kris S. Moe:** Validation, Supervision, Resources, Conceptualization. **Blake Hannaford:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data used in this paper is public and the resources are cited. All code is open-sourced on GitHub with a link provided in the abstract.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT 4 in order to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgment

Blake Hannaford was partially supported by grant #2036255 from the U.S. National Science Foundation.

References

Aganj, I., Harisinghani, M.G., Weissleder, R., Fischl, B., 2018. Unsupervised medical image segmentation based on the local center of mass. Sci. Rep. 8 (1), 13012.

Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.-H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al., 2019. 2017 Robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426.

Angluin, D., 1988. Queries and concept learning. Mach. Learn. 2 (4), 319–342.

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., et al., 2021. Artificial intelligence and machine learning for medical imaging: A technology review. Phys. Medica 83, 242–256.

Belharbi, S., Ben Ayed, I., McCaffrey, L., Granger, E., 2021. Deep active learning for joint classification & segmentation with weak annotator. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3338–3347.

Bodenstedt, S., Rivoir, D., Jenke, A., Wagner, M., Breucha, M., Müller-Stich, B., Mees, S.T., Weitz, J., Speidel, S., 2019. Active learning using deep Bayesian networks for surgical workflow analysis. Int. J. Comput. Assist. Radiol. Surg. 14, 1079–1087.

Bodenstedt, S., Speidel, S., Wagner, M., Chen, J., Kisilenko, A., Müller, B., Maier-Hein, L., Oliveira, B., Hong, S., Zamora-Anaya, J., et al., 2021. HeiChole Surgical Workflow Analysis and Full Scene Segmentation. HeiSurF, MICCAI.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 102062.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision. ICCV.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV.

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. 54, 280–296.

Colleoni, E., Stoyanov, D., 2021. Robotic instrument segmentation with image-to-image translation. IEEE Robot. Autom. Lett. 6 (2), 935–942.

Dougherty, E.R., 1992. An introduction to morphological image processing. In: SPIE. Optical Engineering Press.

Dwibedi, D., Misra, I., Hebert, M., 2017. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1301–1310.

Fuentes-Hurtado, F., Kadkhodamohammadi, A., Flouty, E., Barbarisi, S., Luengo, I., Stoyanov, D., 2019. EasyLabels: weak labels for scene segmentation in laparoscopic videos. Int. J. Comput. Assist. Radiol. Surg. 14, 1247–1257.

Fujita, S., Hagiwara, A., Otsuka, Y., Hori, M., Takei, N., Hwang, K.-P., Irie, R., Andica, C., Kamagata, K., Akashi, T., et al., 2020. Deep learning approach for generating MRA images from 3D quantitative synthetic MRI without additional scans. Invest. Radiol. 55 (4), 249–256.

Funke, I., Jenke, A., Mees, S.T., Weitz, J., Speidel, S., Bodenstedt, S., 2018. Temporal coherence-based self-supervised learning for laparoscopic workflow analysis. In: International Workshop on Computer-Assisted and Robotic Endoscopy. Springer, pp. 85–93.

Gal, Y., Islam, R., Ghahramani, Z., 2017. Deep bayesian active learning with image data. In: International Conference on Machine Learning. PMLR, pp. 1183–1192.

Garcia-Peraza-Herrera, L.C., Fidon, L., D'Ettorre, C., Stoyanov, D., Vercauteren, T., Ourselin, S., 2021. Image compositing for segmentation of surgical tools without manual annotations. IEEE Trans. Med. Imaging 40 (5), 1450–1460.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E.D., Le, Q.V., Zoph, B., 2021. Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2918–2928.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Commun. ACM 63 (11), 139–144.

Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X., 2017. Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hirsch, R., Caron, M., Cohen, R., Livne, A., Shapiro, R., Golany, T., Goldenberg, R., Freedman, D., Rivlin, E., 2023. Self-supervised learning for endoscopic video analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 569–578.

Holub, A., Perona, P., Burl, M.C., 2008. Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. IEEE, pp. 1–8.

Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M., 2011. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Islam, M., Li, Y., Ren, H., 2019. Learning where to look while tracking instruments in robot-assisted surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 412–420.

Kalinin, A.A., Iglovikov, V.I., Rakhlin, A., Shvets, A.A., 2020. Medical image segmentation using deep neural networks with pre-trained encoders. In: Deep Learning Applications. Springer, pp. 39–52.

Kim, T., Lee, K., Ham, S., Park, B., Lee, S., Hong, D., Kim, G.B., Kyung, Y.S., Kim, C.-S., Kim, N., 2020. Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. Sci. Rep. 10 (1), 1–7.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kouw, W.M., Loog, M., 2019. A review of domain adaptation without target labels. IEEE Trans. Pattern Anal. Mach. Intell. 43 (3), 766–785.

Lee, J.H., Choi, I., Kim, M.H., 2016. Laplacian patch-based image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2727–2735.

Lejeune, L., Grossrieder, J., Sznitman, R., 2018. Iterative multi-path tracking for video and volume segmentation with sparse point supervision. Med. Image Anal. 50, 65–81.

Lin, S., Qin, F., Li, Y., Bly, R.A., Moe, K.S., Hannaford, B., 2020. Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 2914–2920.

Lin, S., Qin, F., Peng, H., Bly, R.A., Moe, K.S., Hannaford, B., 2021. Multi-frame feature aggregation for real-time instrument segmentation in endoscopic video. IEEE Robot. Autom. Lett. 6 (4), 6773–6780.

Liu, D., Wei, Y., Jiang, T., Wang, Y., Miao, R., Shan, F., Li, Z., 2020. Unsupervised surgical instrument segmentation via anchor generation and semantic diffusion. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. Springer, pp. 657–667.

Luengo, I., Flouty, E., Giataganas, P., Wisanuvej, P., Nehme, J., Stoyanov, D., 2018. SurReal: Enhancing surgical simulation realism using style transfer. arXiv preprint arXiv:1811.02946.

Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al., 2017. Surgical data science for next-generation interventions. Nat. Biomed. Eng. 1 (9), 691–696.

Peters, B.S., Armijo, P.R., Krause, C., Choudhury, S.A., Oleynikov, D., 2018. Review of emerging surgical robotic technology. Surg. Endosc. 32 (4), 1636–1655.

Qin, F., Li, Y., Su, Y.-H., Xu, D., Hannaford, B., 2019. Surgical instrument segmentation for endoscopic vision with data fusion of cnn prediction and kinematic pose. In: 2019 International Conference on Robotics and Automation. ICRA, IEEE, pp. 9821–9827.

Qin, F., Lin, S., Li, Y., Bly, R.A., Moe, K.S., Hannaford, B., 2020. Towards better surgical instrument segmentation in endoscopic vision: multi-angle feature aggregation and contour supervision. IEEE Robot. Autom. Lett. 5 (4), 6639–6646.

Rajotte, J.-F., Mukherjee, S., Robinson, C., Ortiz, A., West, C., Ferres, J.L., Ng, R.T., 2021. Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary. arXiv preprint arXiv:2101.07235.

Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., et al., 2023. Dissecting self-supervised learning methods for surgical computer vision. Med. Image Anal. 88, 102844.

Raza, K., Singh, N.K., 2021. A tour of unsupervised deep learning for medical image analysis. Curr. Med. Imaging 17 (9), 1059–1077.

Remez, T., Huang, J., Brown, M., 2018. Learning to segment via cut-and-paste. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 37–52.

Roels, J., Saeys, Y., 2019. Cost-efficient segmentation of electron microscopy images using active learning. arXiv preprint arXiv:1911.05548.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al., 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. Int. J. Comput. Assist. Radiol. Surg. 13, 925–933.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. (IJCV) 115 (3), 211–252. http://dx.doi.org/10.1007/s11263-015-0816-y.

Sayari, A.J., Pardo, C., Basques, B.A., Colman, M.W., 2019. Review of robotic-assisted surgery: what the future looks like through a spine oncology lens. Ann. Transl. Med. 7 (10).

Schein, A.I., Ungar, L.H., 2007. Active learning for logistic regression: an evaluation. Mach. Learn. 68 (3), 235–265.

Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N., 2021. A kinematic bottleneck approach for pose regression of flexible surgical instruments directly from images. IEEE Robot. Autom. Lett. 6 (2), 2938–2945.

Sestini, L., Rosa, B., De Momi, E., Ferrigno, G., Padoy, N., 2023. FUN-SIS: A fully unsupervised approach for surgical instrument segmentation. Med. Image Anal. 85, 102751.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Shvets, A.A., Rakhlin, A., Kalinin, A.A., Iglovikov, V.I., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications. ICMLA, IEEE, pp. 624–628.

Singh, N.K., Raza, K., 2021. Medical image generation using generative adversarial networks: A review. In: Health Informatics: A Computational Perspective in Healthcare. Springer, pp. 77–96.

Su, Y.-H., Jiang, W., Chitrakar, D., Huang, K., Peng, H., Hannaford, B., 2021. Local style preservation in improved GAN-driven synthetic image generation for endoscopic tool segmentation. Sensors 21 (15), 5163.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med. Imaging 15 (1), 1–28.

Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J.N., Wu, Z., Ding, X., 2020. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Med. Image Anal. 63, 101693.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P., 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS, IEEE, pp. 23–30.

Tran, T., Do, T.-T., Reid, I., Carneiro, G., 2019. Bayesian generative active deep learning. In: International Conference on Machine Learning. PMLR, pp. 6295–6304.

Wang, T., Lei, Y., Fu, Y., Wynne, J.F., Curran, W.J., Liu, T., Yang, X., 2021. A review on medical imaging synthesis using deep learning and its clinical applications. J. Appl. Clin. Med. Phys. 22 (1), 11–36.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 399–407.

Yoo, T.K., Choi, J.Y., Kim, H.K., 2020. A generative adversarial network approach to predicting postoperative appearance after orbital decompression surgery for thyroid eye disease. Comput. Biol. Med. 118, 103628.

Young, I.T., Van Vliet, L.J., 1995. Recursive implementation of the Gaussian filter. Signal Process. 44 (2), 139–151.

Haonan Peng received his B.Sc. degree in Mechanical Engineering from Central South University, Changsha, China in 2016. In 2019, he received his M.S. degree in mechanical engineering from the University of Washington, Seattle, USA, where he is currently in a Ph.D. program in Electrical and Computer engineering. His research interests include surgical robotics, machine learning, and haptic feedback.

Shan Lin received the B.Sc. degree in electrical engineering from Xiamen University, Xiamen, China, in 2015, and her M.S. degree in electrical engineering from Vanderbilt University, Nashville, USA, in 2017, and her Ph.D. degree in electrical and computer engineering from the University of Washington, Seattle, USA, in 2021. She is currently a Postdoctoral Researcher at the University of California, San Diego. Her research interests include robot vision, surgical robotics, and deep learning.

Daniel King received his B.Sc. degree in Mechanical Engineering from Purdue University in 2014 and his M.S. degree in electrical and computer engineering from the University of Washington in 2021. His research interests include robotics, machine learning, and computer vision.

Yun-Hsuan (Melody) Su received her Ph.D. degree in Electrical and Computer Engineering at the University of Washington, with a focus on medical robotics in 2020. In 2018, Dr. Su was a Research Engineer and worked on visual and force servoing for industrial robots at ABB robotics, leading to two patent applications. She is passionate about outreach STEM programs and has been closely involved in IEEE TryEngineering, and the Pioneer Academics Program. Currently, she is an Assistant Professor in Computer Science at Mount Holyoke College. Dr. Su's research interests include surgical robotics, vision-based force estimation, computer/machine vision, and haptic feedback.

Waleed M. Abuzeid, M.D., is an Associate Professor and fellowship-trained rhinology/skull base surgery specialist in the Department of Otolaryngology - Head and Neck Surgery at the University of Washington. Dr. Abuzeid received his B.Sc. in Physiology from University College London in 2002 and his M.B.B.S. in Clinical Practice from University College London Medical School in 2005. Dr. Abuzeid works alongside partners in UW Engineering and the tech industry to innovate and develop new technologies such as, for example, artificial intelligence supported clinical decision making, real-time surgical navigation, and augmented reality.

Randall A. Bly is an Associate Professor with the Department of Otolaryngology - Head and Neck Surgery at the University of Washington School of Medicine. His clinical practice is at Seattle Children's Hospital where he cares for patients and families with a variety of problems in the head and neck. He has specific interests in patients with speech and swallowing problems, vascular malformations, breathing difficulties, neck masses, and congenital anomalies. His research interest is in how surgery is planned, guided, and analyzed in order to improve patient outcomes, reduce the length of hospital stays, and recover more quickly.

Kris. S. Moe, MD, FACS, is Chief of Facial Plastic and Reconstructive Surgery and a Professor in the Departments of Otolaryngology – Head & Neck Surgery and Neurological Surgery at the University of Washington in Seattle. Among his major clinical and research efforts has been the development of transorbital neuroendoscopic surgery (TONES), a minimally disruptive scarless technique for operating on the skull base and brain. He has received significant research grant funding from multiple sources including the NIH. He is an innovator of medical devices, holds multiple patents, and has published well over 100 peer-reviewed articles, book chapters and books.

Blake Hannaford received the B.S. degree in Engineering and Applied Science from Yale University in 1977, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of California, Berkeley. From 1986 to 1989 he worked on the remote control of robot manipulators at NASA Jet Propulsion Laboratory, Caltech and supervised that group from 1988 to 1989. Since 1989, he has been at the University of Washington, where he is Professor of Electrical Engineering. He was at Google-X / Verily from 2014 to 2015.