SurgicalSemiSeg: A Semi-Supervised Framework for Laparoscopic Image Segmentation

Yuning Zhou[1] YUNIZHOU@STUDENT.UNIMELB.EDU.AU
Henry Badgery[2] HENRY.BADGERY@SVHA.ORG.AU
Matthew Read[3] MATTHEW.READ@UNIMELB.EDU.AU
James Bailey[4] BAILEYJ@UNIMELB.EDU.AU
Catherine E. Davey[1] CATHERINE.DAVEY@UNIMELB.EDU.AU
[1] *Department of Biomedical Engineering, the University of Melbourne, Australia*
[2] *Department of HPB/UGI Surgery, St Vincent's Hospital Melbourne, Australia*
[3] *Department of Surgery, St Vincent's Hospital Melbourne, Australia*
[4] *School of Computing and Information Systems, the University of Melbourne, Australia*

Editors: Under Review for MIDL 2025

Abstract

Deep learning applications in surgery are heavily reliant on large-scale datasets with high-quality annotations, which are costly and time-consuming to obtain. Self-supervised learning (SSL) has shown significant potential for reducing reliance on labelled data. This work investigates the use of SSL for semantic segmentation in laparoscopic cholecystectomy (LC) surgery. Through evaluation of existing SSL methods, we find that pixel-level objectives enable the most effective representation learning for laparoscopic imaging, characterised by highly variable and deformable anatomy. Building on this insight, we develop a tailored masked denoising autoencoder with a carefully optimised masking ratio and patch size for semantic segmentation. Our method achieves state-of-the-art performance across three LC datasets. Of note, it significantly improves segmentation accuracy for critical anatomical structures that are under-represented in training datasets. Furthermore, our approach achieves generalisability, with pre-trained representations performing effectively across fine-tuning datasets from different LC datasets.

Keywords: Self-supervised learning, laparoscopic imaging, semantic segmentation

1. Introduction

Deep learning-based precise surgical scene interpretation, such as semantic segmentation, is a crucial component of AI based intraoperative guidance tools designed to enhance surgical safety. The training of deep neural networks (DNNs) for semantic segmentation requires large-scale datasets with meticulous pixel-level annotations, that are costly and labour-intensive to produce. The development of medical image segmentation datasets includes two major challenges: i) *significant variations in the appearance of anatomical structures and surgical instruments*, and ii) *class imbalance in under-represented structures*. These challenges have impaired the accuracy of surgical image neural networks, limiting the potential for real world clinical application (Tokuyasu et al.; Maqbool et al., 2020; Silva et al., 2022; Yoon et al., 2022).

Recently, self-supervised learning (SSL) approaches have been employed in surgical computer vision applications to leverage high volume unlabelled data to enhance the performance of DNNs, mitigating the challenges of developing sufficiently large annotated datasets. SSL involves training models on carefully designed pretext tasks using unlabelled data. This pre-trained model can then be fine-tuned on downstream tasks, progressively improving the performance compared to simply training a model on labelled datasets (Chen et al., 2020). While SSL has been employed in the literature for classification of structures on surgical images and videos (Kletz et al., 2019b; Twinanda et al., 2016; Jin et al., 2018; Mishra et al., 2017; Hashimoto et al., 2019; Kitaguchi et al., 2020), its use for segmentation of anatomy and instruments in surgery has not been widely investigated.

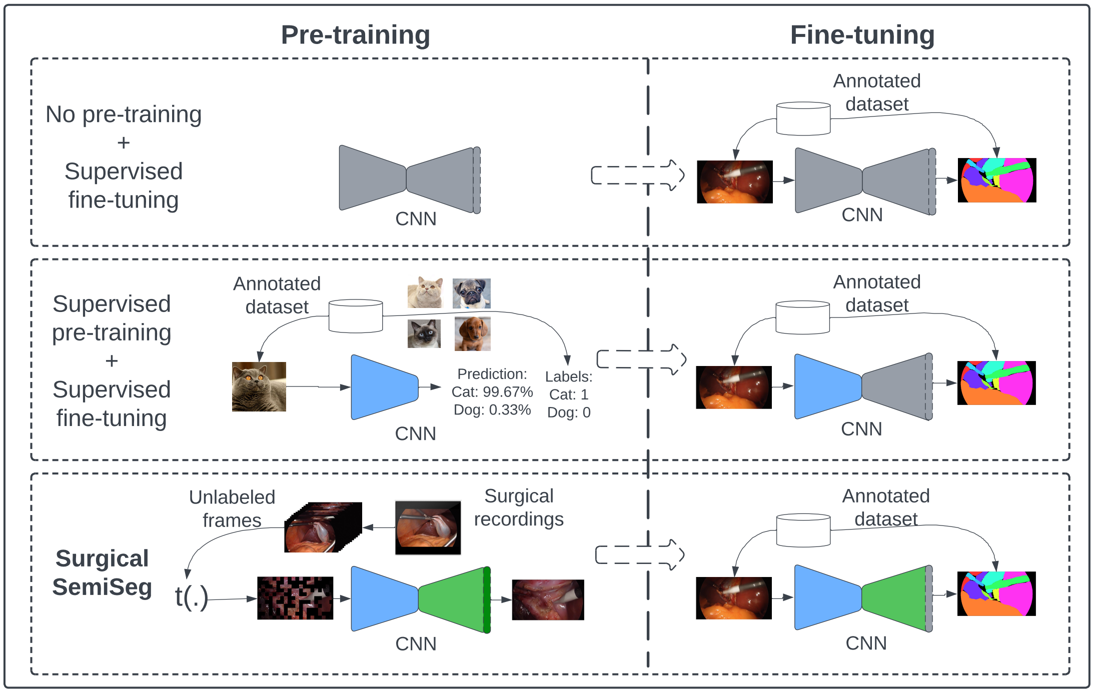


Figure 1: Illustration of two-stage training frameworks for semantic segmentation with three types of pre-training strategies. From top to bottom: no pre-training, supervised pre-training, and our Surgical Semi-supervised Segmentation (SurgicalSemiSeg) framework with tailored denoising autoencoder designs as pre-training. CNNs colours indicate re-used parameters from pre-training.

In this paper, we evaluate common pretext tasks for static images such as random rotation (Gidaris et al., 2018), colourisation (Zhang et al., 2016), autoencoder (Hinton and Salakhutdinov, 2006), and denoising autoencoder (Vincent et al., 2008), alongside advanced methods like contrastive learning (SimCLR) (Chen et al., 2020), masked autoencoder (MAE) (He et al., 2022), and a recent contrastive method tailored for LC segmentation (DDA) (Zhou et al., 2024). Through extensive evaluation, we observe that pixel-level generation tasks are effective for segmentation due to their alignment with pixel-level objectives. Building on this observation, we propose our Surgical Semi-supervised Segmentation framework (SurgicalSemiSeg). Figure 1 demonstrates the framekwork schematic. This framework involves a denoising autoencoder for self-supervised pre-training, and a supervised fine-tuning step. Specifically, we introduce four masking parameters for the denoising autoencoder, each for a distinct input corruption strategy. These parameters provide flexibility in the masking process, allowing it to operate independently of specific token positions as in a masked autoencoder. Additionally, unlike image-level pre-training approaches that disregard the decoder during fine-tuning (Chen et al., 2020), the pixel-level pre-training objective allows SurgicalSemiSeg to fully preserve the pre-trained model while modifying only the decoder's final layer during fine-tuning. This approach ensures that the understanding capability gained by both the encoder and decoder during pre-training is largely retained, maximising the utilisation of unlabelled data to improve segmentation performance in the downstream task.

In summary, the contributions of this paper are as follows:

- We identify that self-supervised objectives at the pixel level are the most effective for segmentation tasks in surgical contexts.
- We present a two-stage Surgical Semi-supervised Segmentation framework (SurgicalSemiSeg). It allows easy plug-in and play with different datasets and deep learning architectures, while maximising the preservation of pre-trained representations.
- We propose a tailored denoising autoencoder with optimal mask designs as a pre-training objective in SurgicalSemiSeg, which significantly improves the segmentation accuracy of under-represented but clinically important classes.
- SurgicalSemiSeg outperforms other baseline SSL methods across three downstream datasets and demonstrates robust and transferable representations across different institutions.

2. Preliminaries

Given a dataset, $\mathcal{D}$, comprised of an unlabelled subset for pre-training, $\mathcal{D}_u$, and a labelled subset for fine-tuning, $\mathcal{D}_l$, we define $\mathcal{D} = \mathcal{D}_u \cup \mathcal{D}_l$, such that $\{\boldsymbol{x}_i\}_{i=1}^{p} \in \mathcal{D}_u$ and $\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1}^{q} \in \mathcal{D}_l$, with p, q denoting the samples sizes in the unlabelled and labelled subsets respectively. In a typical scenario, $p \gg q$. $\boldsymbol{x} \in \mathbb{R}^{w \times h \times 3}$ denotes an input image of width w and height h in the red-green-blue (RGB) space, and $\boldsymbol{y}$ denotes the corresponding (pixel) labels.

A DNN model, $f_\theta = h \circ g$, is assumed to be a CNN with an encoder-decoder architecture, parameterized by θ. The encoder, h, maps input, $\boldsymbol{x}$, to a set of deep (or latent) features in the high ($C' \gg 3$) dimensional space, $\boldsymbol{z} = h(\boldsymbol{x})$, where $\boldsymbol{z} \in \mathbb{R}^{w' \times h' \times C'}$. The input spatial resolution gradually decreases while the number of channels (features) increases when passing an input through multiple convolutional layers in the encoder. Thus, $w' \ll w, h' \ll h$. The decoder, g, generates $\boldsymbol{z}$ to the desired output according to the objective and makes the final predictions via its last layer.

In supervised learning, DNN training can be formulated as the following optimisation problem:

$$\arg\min_{\theta} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_l} \mathcal{L}(f_\theta(\boldsymbol{x}), \boldsymbol{y}), \tag{1}$$

where $\mathcal{L}$ is the objective function (e.g., cross-entropy) calculating the difference between model prediction and the ground truth label.

In self-supervised learning, the ground truth is generated from the input data $\boldsymbol{x}$, here noted as $\boldsymbol{x}'$. In this case, self-supervised training can be formulated as the following optimisation problem:

$$\arg\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}_l}\mathcal{L}(f_\theta(\boldsymbol{x}), \boldsymbol{x}'), \tag{2}$$

where $\mathcal{L}$ is the objective function (e.g., Euclidean distance) calculating the difference between model prediction and the generated ground truth.

3. Surgical semi-supervised framework

We provide an overview of our surgical semi-supervised framework in Section 3.1, and propose the masked-corrupted denoising autoencoder pre-training objective in Section 3.2.

3.1. Two-stage framework

We present a simple yet versatile two-stage semi-supervised learning framework, name SurgicalSemiSeg, designed to maximally exploit unlabelled surgical images to improve segmentation performance. Figure 1 illustrates the difference of three types of training frameworks for semantic segmentation, including no pre-training, supervised pre-training, and the proposed SurgicalSemiSeg with a tailored mask-corrupted denoising autoencoder as the pre-training objective, which will be explained in Section 3.2. With the pixel-level objectives in both stages, our framework allows flexible integration of any segmentation model with encoder-decoder architecture, making it highly adaptable for different applications. For this study, we adopt a CNN for its flexibility on different input resolutions.

Additionally, SurgicalSemiSeg allows the maximal preservation of self-supervised pre-trained representations by reinitialising the final layer weights during fine-tuning. To assess pre-trained representation quality, pre-trained models require fine-tuning on a target annotated dataset. This process transfers the representations learned during the first stage (pre-training on large-scale unlabelled surgical video frames, here we only focus on static image pretext tasks in the scope of our paper) to a smaller-scale annotated segmentation dataset. Existing image-level pre-training (Hinton and Salakhutdinov, 2006; Gidaris et al., 2018; Chen et al., 2020) usually re-uses only encoder weights in fine-tuning. Our framework benefits from the semantic understanding and spatial reconstruction ability from the pre-trained decoder by modifying its last layer only. This design ensures a seamless transition from pre-training to fine-tuning, requiring adjustments only to the dataset and learning objective, while maintaining the integrity of the pre-trained representations.

3.2. Mask-corrupted denoising objective

Segmentation of surgical images poses unique challenges, including i) *extreme class imbalance due to varying class size, appearance and occurrence*, ii) *hard-to-delineate objects with overlap* (Ferguson et al., 1992; Asbun et al., 1993), iii) *predominantly reddish content in both the background and foreground of surgical images*, and iv) *significant variation in perspective due laparoscope movement, lighting variation and difference operative aproach.*

Conventional denoising autoencoders apply Gaussian noise to the input pixels, with the model inferring the corrupted pixels based solely on the surrounding pixel values. In surgical views, the challenges mentioned above cause nearby pixel values within the same frame to be highly similar. This similarity restricts the quality of representation learning, as it becomes difficult for the model to distinguish between anatomies that share similar color characteristics. As a result, this approach may fail to capture generalised and meaningful representations for segmentation tasks.

To address these challenges and learn generalised representations, we propose a specially designed mask-corrupted denoising autoencoder tailored specifically for surgical segmentation. Inspired by He et al. (2022), which has demonstrated exceptional representation learning capabilities by generating large portions of missing input patches, we hypothesise that denoising autoencoders can similarly benefit from patch-based noise. By corrupting larger regions of the image, reconstruction becomes more challenging, since the model requires to infer missing parts of structures or even entire objects from masked-out areas.

To facilitate accurate pixel-class predictions, the pre-trained model can benefit from recognising pixels within objects and differentiate them from those between objects. We propose a novel mask design with four parameters to enable the application of patch-based masks with varying sizes and masking ratios across the entire image. This mask-design noise enables the model to learn generalised representations that remain consistent despite variation in object appearance, occlusions, perspective changes, thus addressing the unique challenges of surgical image segmentation.

Given an input image $\boldsymbol{x} \in \mathbb{R}^{w \times h \times 3}$ and a binary mask $\boldsymbol{m} \in \{0,1\}^{w \times h}$, a masked transformation function, $t_{\{\rho,N,s,c\}}$ with four parameters on the input image, is defined as $\boldsymbol{x}' = t(\boldsymbol{x}) = \boldsymbol{x} \odot \boldsymbol{m} + (1 - \boldsymbol{m}) \odot \delta$, where $\odot$ is element-wise multiplication applied to each RGB channel, and δ has matching dimensions with $\boldsymbol{x}$ and contains the replacement value for each masked pixel (default is 0). Four masking parameters are described as follows:

- $\rho \in [0\%, 90\%]$**:** the ratio of masked pixels, or pixels with 0 values in $\boldsymbol{m}$, among the total pixels in the input. $\rho = 0\%$ simplifies the masked pre-training to an autoencoder.
- $N \in [8, 256]$**:** the side length of an individual square mask patch or the diameter length of a circle mask.
- s**:** the mask component shape. For simplicity we focus on square and circle masks.
- c**:** the replacing value in δ, also known as mask colour. We adopted black or random colours in the masks for every pixel or mask component.

An illustration of different masking parameters is provoded in Figure 2. The mask-corrupted input $t(\boldsymbol{x})$ and the clean input $\boldsymbol{x}$ as input-reference pairs are then input into a CNN with encoder-decoder architecture. The encoder extracts deep representations of the whole $t(\boldsymbol{x})$ with mask corruption as $\boldsymbol{z}' = h(t(\boldsymbol{x}))$. The decoder g then transforms the deep representations of the unmasked input regions (entangled with mask-corrupted noise) back into the input space, as $g(\boldsymbol{z}') \in \mathbb{R}^{w \times h \times 3}$. We adopt the optimisation objective below to minimise the pixel-wise reconstruction differences:

$$\arg\min_{\theta} \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_u} \frac{1}{h \times w} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \left\| f_\theta(t(\boldsymbol{x}))_{ij} - \boldsymbol{x}_{ij} \right\|^2 . \tag{3}$$

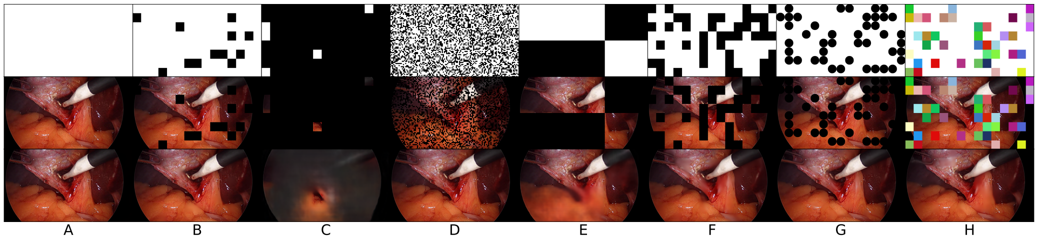


Figure 2: Illustration of mask design parameters on a single example, from top to bottom row, shows the masks, masked images, and reconstructed images under different mask settings (ρ for masking ratio and N for mask size): A. no mask, B. $\rho = 10\%, N = 64$, C. $\rho = 90\%, N = 64$, D. $\rho = 40\%, N = 8$, E. $\rho = 40\%, N = 256$, F. $\rho = 40\%, N = 64$ (optimal mask settings), G. $\rho = 40\%, N = 64$ in circle masks, H. $\rho = 40\%, N = 64$ in coloured masks.

SurgicalSemiSeg maximises preservation of the *entire* pre-trained model developed on the unlabelled dataset. Unlike image-level pretext tasks that lack a decoder, our method includes a pre-trained decoder that has learned to resolve semantic and spatial relationships between pixels from the encoder representation during pre-training. This design ensures that the semantic and spatial knowledge acquired in pre-training is largely retained during fine-tuning. By preserving meaningful structural and contextual information, SurgicalSemiSeg with the masked-corrupted denoising autoencoder fully exploits the potential of pixel-level self-supervised representation learning for segmentation tasks.

4. Experiments

4.1. Experiment settings

DeepLabV3+ (Chen et al., 2018) with ResNet101 (He et al., 2016) backbone is adopted as the default model. For each pre-training, the model was trained for 20 epochs with 16 as batch size, AdamW (Loshchilov and Hutter, 2019) as the optimiser, 0.001 as learning rate, and 0.01 weight decay. All fine-tuning applies the same parameters, except changing the learning rate to 0.005. For computational efficiency, we resized the in-house images to 960 × 540 and followed the original resolutions for public datasets. Augmentations of 10 degrees of rotation, horizontal flipping, and colour jittering (with brightness 0.25, contrast 0.25, saturation 0.25, and hue 0.0) applied. Experiments were conducted on 4 A100 GPUs with PyTorch implementations.

As mentioned datasets can be easily plugged-in and play in our framework, therefore, we validate the transferability of representation learned between different institutions. We used an in-house LC dataset as well as a public LC datasets (Twinanda et al., 2016) for pre-training, and validated the pre-training methods on our in-house dataset and two public datasets (Hong et al., 2020; Maqbool et al., 2020). The dataset descriptions are explained in Appendix A.

4.2. Comparison with existing pre-text tasks

We evaluate the effectiveness of different pre-training strategies on under-represented classes across three datasets. For our mask-corrupted denoising autoencoder, we adopt the optimal

Table 1: Performance of different pre-training strategies on three validation datasets. *IoU* is reported in percentage. The best results are in **bold**.

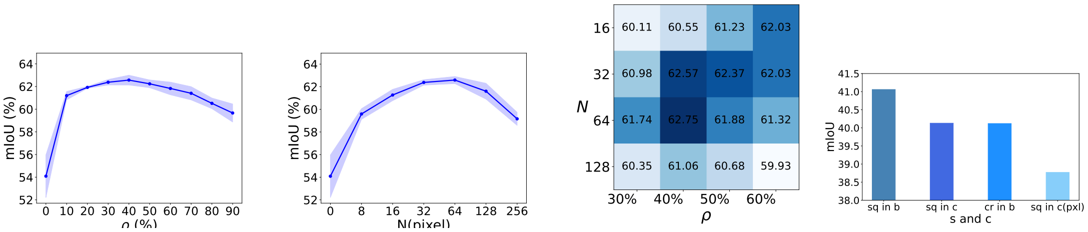
Datasets	All	Under represented	Number	Pre-training Strategies								
				N/A	Supervised	Rotation	Colourisation	Autoencoder	SimCLR	MAE	DDA	Ours
M2caiSeg	✓		19	68.37	80.46	73.87	79.60	76.32	81.79	72.04	**85.37**	82.38
		✓	12	56.40	72.07	63.67	70.55	66.37	73.76	61.81	**78.60**	74.55
CholecSeg8k	✓		8	57.49	61.59	55.33	64.52	56.71	56.35	57.94	61.71	**66.05**
		✓	1	1.85	33.29	44.50	28.33	22.64	35.48	33.29	40.98	**59.15**
In-house Seg	✓		20	56.03	59.16	56.20	59.15	50.37	58.57	61.63	58.44	**62.26**
		✓	11	42.23	45.94	43.01	45.92	37.79	45.10	44.73	48.72	**50.67**

Table 2: Class-wise performance of different pre-training strategies on In-house Seg. *IoU* is reported in percentage. The best results are in **bold**.

	Class Names	Pre-training Strategies								
		N/A	Supervised	Rotation	Colourisation	Autoencoder	SimCLR	MAE	DDA	Ours
N/A	background	97.56	97.87	97.36	97.87	89.67	98.00	**98.08**	97.68	97.96
Instruments	cholangiogram catheter	63.76	74.80	69.02	71.16	61.24	76.36	75.86	73.50	**76.97**
	clip applicator	49.96	**60.73**	45.39	53.78	39.10	48.68	60.52	47.72	58.01
	diathermy hook shaft	71.52	73.54	71.75	72.49	64.22	72.13	**81.39**	73.36	73.69
	diathermy hook tip	81.59	84.15	83.39	84.49	80.33	84.12	**87.17**	81.73	85.19
	grasper shaft	76.34	78.89	76.00	79.06	64.86	79.21	**80.35**	78.88	80.21
	grasper tip	57.50	64.38	54.70	64.00	53.70	65.18	64.35	61.61	**67.24**
	scissors shaft	7.60	5.93	5.20	1.54	7.70	6.06	**19.11**	11.66	12.27
	scissors tip	25.36	27.67	36.11	27.75	34.03	24.33	**59.47**	23.91	45.27
	sucker irrigator	55.76	63.38	52.53	62.21	45.06	62.83	64.43	63.27	**65.63**
Anatomies	abdomen wall	36.16	41.44	34.71	41.20	12.81	39.23	39.73	**42.55**	41.06
	common bile duct	55.59	50.95	57.60	**59.67**	45.35	56.44	59.51	56.76	56.89
	cystic artery	27.58	25.08	27.86	32.51	24.94	28.65	**33.75**	25.46	33.20
	cystic duct	49.79	50.46	48.78	50.38	46.68	50.36	51.13	51.08	**52.96**
	duodenum	13.75	25.69	18.75	24.86	17.32	25.11	20.54	23.46	**28.38**
	gallbladder	80.26	81.85	79.95	82.52	78.19	**82.76**	81.78	81.40	82.34
	liver	84.96	87.01	84.58	86.50	79.61	85.82	85.50	85.53	**87.79**
	omentum	87.49	87.53	87.76	88.61	72.09	**88.88**	88.96	88.55	88.82
	rouviere's sulcus	17.26	18.70	12.90	20.10	11.95	16.35	0.00	11.83	**26.85**
	segment iv	80.89	83.26	79.64	82.29	78.57	80.88	81.02	83.84	**84.38**
Mean		56.03	59.16	56.20	59.15	50.37	58.57	61.63	58.44	**62.26**

mask parameters (the searching process is illustrated in Figure 3, and further described in Section B) searched under In-house datasets as default settings. The results, summarised in Table 1, show the average performance of all classes and specifically for under-represented classes, defined as those comprising less than 1% of the pixel distribution. Except for M2caiSeg, which is a very small dataset, pixel-level pretext tasks generally outperform image-level ones. Our method notably improves prediction accuracy, especially for under-represented classes.

We further report the class-wise IoU of different pre-training methods on In-house Seg in Table 2. For critical anatomical structures foreratiive safety, such as the common bile duct and omentum, our method improves the baselines by 5.27% and 4.36%. It also significantly improves the recognition of scissors (7.61%), a challenging class easily confused with other instruments (Kletz et al., 2019a; Jaafari et al., 2021; Namazi et al., 2022).



N	30%	40%	50%	60%
16	60.11	60.55	61.23	62.03
32	60.98	62.57	62.37	62.03
64	61.74	62.75	61.88	61.32
128	60.35	61.06	60.68	59.93

(a) masking ratio ρ (b) patch size N (c) grid search (d) shape and colour

Figure 3: Influence of varying mask parameters on In-house Seg (in mIoU). Results are reported as mIoU (in percentage). Curves with filling show the mean and standard deviation over 5 random seeds. Darker colour in (c) indicated better performance.

Table 3: Comparison of different pre-training dataset on three downstream datasets. *IoU* is reported in percentage. The best results are in **bold**.

Fine-tuning Datasets	Pre-training Datasets	
	Inhouse Unlabelled	Cholec80
M2caiSeg	82.38	**83.31**
Cholecseg8k	66.05	**68.23**
In-house Seg	**62.26**	60.71

4.3. Generalised representations across institutions

Table 3 demonstrates that our method achieves the best performance when the pre-training and fine-tuning datasets are collected from the same institution, where there is less variation between surgical equipments and operative techniques. Furthermore, the results indicate that representation learning from similar operations, in this case, LC, generalise well across different institutional datasets. This finding highlights the potential and effectiveness of leveraging unlabelled surgical recordings to enhance deep learning applications in surgery. While our results specifically validate the method in LC, the approach is likely to perform well across other surgical procedures. Our pre-trained models on Cholec80 will be made publicly available.

5. Conclusion

In this paper, we conduct an extensive evaluation of self-supervised learning on static image for LC segmentation. Based on our findings that aligned objectives of pre-training and fine-tuning enable the most effective representation learning, we propose SurgicalSemiSeg, a semi-supervised framework with a tailored masked denoising autoencoder for laparoscopic images and provide comprehensive design guidelines. Our method significantly enhances the recognition of under-represented classes that are safety related. This simple yet powerful method offers valuable insights into leveraging unlabelled data for computer-assisted surgery applications. Furthermore, our generalisable and open-sourced pre-trained model serves as a valuable resource for the community, facilitating the development of LC segmentation applications.

References

Horacio J Asbun, Ricardo L Rossi, Jeffrey A Lowell, and J Lawrence Munson. Bile duct injury during laparoscopic cholecystectomy: mechanism of injury, prevention, and management. *World journal of surgery*, 17:547–551, 1993.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *NeurIPS*, 2020.

Charles M Ferguson, David W Rattner, and Andrew L Washaw. Bile duct injury in laparoscopic cholecystectomy. *Surgical Laparoscopy Endoscopy & Percutaneous Techniques*, 2 (1):1–7, 1992.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations, 2018.

Daniel A Hashimoto, Guy Rosman, Elan R Witkowski, Caitlin Stafford, Allison J Navarette-Welton, David W Rattner, Keith D Lillemoe, Daniela L Rus, and Ozanan R Meireles. Computer vision analysis of intraoperative video: automated recognition of operative steps in laparoscopic sleeve gastrectomy. *Annals of surgery*, 270(3):414–421, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

W.-Y. Hong, C.-L. Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. 2020.

Jaafar Jaafari, Samira Douzi, Khadija Douzi, and Badr Hssina. Towards more efficient cnn-based surgical tools classification using transfer learning. *Journal of Big Data*, 8: 1–15, 2021.

Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 691–699. IEEE, 2018.

Daichi Kitaguchi, Nobuyoshi Takeshita, Hiroki Matsuzaki, Hiroaki Takano, Yohei Owada, Tsuyoshi Enomoto, Tatsuya Oda, Hirohisa Miura, Takahiro Yamanashi, Masahiko Watanabe, et al. Real-time automatic surgical phase recognition in laparoscopic sigmoidectomy using the convolutional neural network-based deep learning approach. *Surgical endoscopy*, 34:4924–4931, 2020.

Sabrina Kletz, Klaus Schoeffmann, Jenny Benois-Pineau, and Heinrich Husslein. Identifying surgical instruments in laparoscopy using deep learning instance segmentation. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, 2019a.

Sabrina Kletz, Klaus Schoeffmann, and Heinrich Husslein. Learning the representation of instrument images in laparoscopy videos. *Healthcare Technology Letters*, 2019b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks, 2020.

Kaustuv Mishra, Rachana Sathish, and Debdoot Sheet. Learning latent temporal connectionism of deep residual visual abstractions for identifying surgical tools in laparoscopy procedures. In *CVPR*, 2017.

Babak Namazi, Ganesh Sankaranarayanan, and Venkat Devarajan. A contextual detector of surgical tools in laparoscopic videos using deep learning. *Surgical endoscopy*, pages 1–10, 2022.

Bruno Silva, Bruno Oliveira, Pedro Morais, LR Buschle, Jorge Correia-Pinto, Estevão Lima, and Joao L Vilaça. Analysis of current deep learning networks for semantic segmentation of anatomical structures in laparoscopic surgery. In *EMBC*, 2022.

Tatsushi Tokuyasu, Yukio Iwashita, Yusuke Matsunobu, Toshiya Kamiyama, Makoto Ishikake, Seiichiro Sakaguchi, Kohei Ebe, Kazuhiro Tada, Yuichi Endo, Tsuyoshi Etoh, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surgical endoscopy*.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 2016.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Jihun Yoon, SeulGi Hong, Seungbum Hong, Jiwon Lee, Soyeon Shin, Bokyung Park, Nakjun Sung, Hayeong Yu, Sungjae Kim, SungHyun Park, et al. Surgical scene segmentation using semantic image synthesis with a virtual surgery environment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 551–561, 2022.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016.

Yuning Zhou, Henry Badgery, Matthew Read, James Bailey, and Catherine Davey. DDA: Dimensionality driven augmentation search for contrastive learning in laparoscopic surgery. In *Medical Imaging with Deep Learning*, 2024.

Appendix A. Datasets description

The 5 datasets are comprehensively described in this section, including two unlabelled datasets for pre-training and three labelled datasets for fine-tuning. The collection process are the same with (Zhou et al., 2024)

The two unlabelled datasets adopted for elf-supervised pre-training are described as follows:

- **In-house Unlabelled** contains 300,000 frames at 4 fps from the other 50 videos to avoid data leakage. It is the default unlabelled dataset used for self-supervised pre-training.

- **Cholec80 (Unlabelled)**(Twinanda et al., 2016) is a publicly available classification dataset with tool presence and surgical phase annotations. It contains 80 videos in 480 × 854, where each video denotes an individual patient. We only make use of the frames for self-supervised pre-training and disregard the annotations. To avoid data leakage, we also discard the mutual videos in CholecSeg8k and m2caiSeg, which are two subsets of Cholec80 with semantic segmentation annotations. In this way, the adopted unlabeled Cholec80 dataset ends up with 63 videos, which generated 400,000 frames at 2 fps.

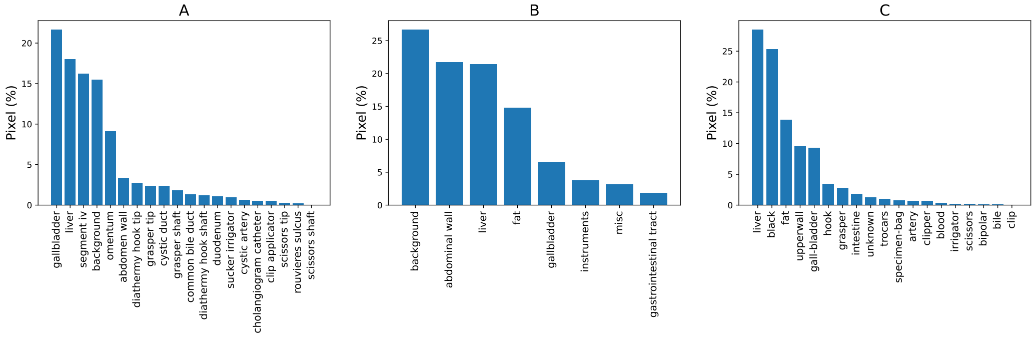


Figure 4: Class distribution of pixels in three LC segmentation datasets, A:In-house Seg B:CholecSeg8k C:M2caiSeg.

The three labelled datasets and their usage are described as below, with the statistics of their class distributions are displayed in Figure 6.

- **In-house Seg** contains frames selected from 20 videos. The individual frames in in 32fps target clips were first pulled out, following by a pixel-wise threshold selection which compares the concecutive frames with the anchor frame and select the next frame if only its pixel difference exceed the threshold. This yielded 4,136 frames in total, where the training set contains 3,740 frames from 16 videos, and the test set contains 392 frames from 4 videos that are unseen in the training set. The dataset was annotated and validated by our surgeons. To evaluate the pre-training strategies and DNNs structure recognition effectiveness in the real-world surgical context, we explicitly defined the semantic class and include *9 surgical instruments* and *10 anatomical* *structures* shown up during the interested surgical phases. The fine-grained class definition cause the data distribution extremely skewed, which well-represented the real-world challenge.

- **CholecSeg8k** (Hong et al., 2020) is a labeled subset of Cholec80 which contains 8,080 frames of 480 × 854 at 25 fps from 17 videos in Cholec80. Following (Silva et al., 2022), we merge the 13 semantic classes into 8 classes under the same train-test split.
- **M2caiSeg** (Maqbool et al., 2020) is a labelled subset of the MICCAI 2016 Surgical Tool Detection dataset (Twinanda et al., 2016), which contains videos from Cholec80. It contains 307 frames in 596 × 334 from 2 videos annotated with 19 classes.

Table 4: Class description and their colour encoding in R, G, B of In-house Seg.

Class name	Description	Colour code (RGB)
abdominal wall	abdominal wall	100, 20, 80
background	black background beyond circular visual field	200, 200, 200
cholangiogram catheter	instrument to apply dye-enhanced imaging for bile ducts visulization (includes shaft, trip and catheter)	0, 130, 170
clip applicator	instrument to apply clips to close cystic artery and duct (includes shaft, trip and catheter)	130, 130, 0
common bile duct	bile duct drain from hepatic ducts to duodenum	0, 250, 200
cystic artery	blood supply to the gallbladder	255, 255, 0
cystic duct	duct draining bile from gallbladder to common bile duct	64, 255, 50
diathermy hook shaft	diathermy hook instrument - shaft	49,249,166
diathermy hook tip	diathermy hook instrument - tip	0, 190, 80
duodenum	dection of gastrointestinal tract where common bile duct drains, distal to stomach	20, 102, 73
gallbladder	gallbladder	50, 255, 255
grasper shaft	grasping instrument of any kind - shaft	50, 193, 255
grasper tip	grasping instrument of any kind - tip	50, 132, 255
liver	all other liver segments	255, 0, 0
omentum	intra-abdominal fat, includes small bowel	255, 197, 50
rouviere's sulcus	cleft on the right side of the liver; important landmark	255, 182, 193
scissors shaft	instrument to cut tissues and structures	180, 50, 255
scissors tip	instrument to cut tissues and structures	214, 50, 255
segment iv	segment of liver to the patient left side of gallbladder	165, 42, 42
sucker irrigator	cylindrical instrument for suction and irrigation	100, 0, 130

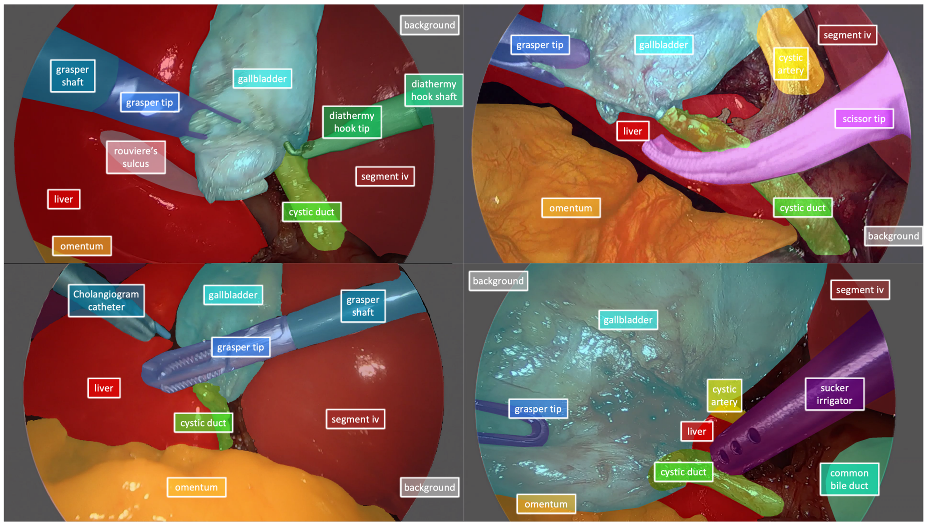


Figure 5: Examples of annotated frames in In-house Seg overlaid with the class colour mask

Appendix B. Masked denoising autoencoder designs

To reduce computation cost, we first search for the optimal masking ratio with a fixed N, and then fix the optimal ρ and investigate N. For in-house Seg, the optimal performance is reached when masking 40% with fixed N to 64, and the performance is relatively stable between 30% and 60%. Varying N has a more significant influence on downstream performance, which peaked at $N = 64$ when $\rho = 40\%$. We further conducted a grid search with ρ and N in a similar performance range and confirmed that the optimal setting for in-house Seg is $\rho = 40\%$ and $N = 64$ in Figure 3(c). On this optimal setting, results in Figure 3(d) show the black square mask is preferable. Unless explicitly stated, we use $\rho = 40\%$ and $N = 64$ with the black square mask as the default setting in following sections for In-house Seg based on this observation.

For the two public datasets, increasing masking ratio and patch size also results in upward parabola in segmentation performance. m2caiSeg is less sensitive to ratio changes, but demonstrates higher performance variance across different runnings of the models under the same mask settings, which is a common challenge in deep learning with small dataset. On CholecSeg8k, the optimal ρ is observed at 20% and N at 32.

Appendix C. Dataset granularity

In this section, we explore the effectiveness of our pre-training on the same dataset with different class definitions. Due to the dataset annotation difficulty, it is infeasible for institutions to extensivelly labelled everything appeared during the surgical procudure. Based on different clinical focus, we defined our extensively labelled dataset with 4 different number and granularity of class difinitions.

- *Only Instrument* includes 6 commonly seen instruments in the duration of procedure we are focusing on, including the cholangiogram catheter, clip applicator, diathermy hook, grasper, scissors and sucker irrigator. Since it is common in robotic surgery dataset that the surgical tool tips and shafts are annotated seperately for precise tool motion tracking, we seperate the shaft and tip for three of the most frequently occured instruments that could possibly cause tissue or organ damage, the diathermy hook, grasper and scissors.

- *Only Critical Anatomies* includes 5 anatomical structures related mainly to the dissection, namely cystic artery, cystic duct, duodenum, gallbladder, and rouvieres sulcus. Within them, cystic artery, cystic duct are the two critical structures for dissection. Duodenum is regarded as the danger zone in dissectoin located below cystic artery, cystic duct. It should be never approached to during the dissection. Rouvieres sulcus is an rarely appeared landmark structure. With its occurrence, the surgical instruments for dissection, like the diathermy hook or scissors, should never operate on any anatomical structures below it.

- *Only Easy Anatomies* includes 4 anatomies, duodenum, gallbladder, liver, and segment iv. They have relatively larger volume compared to other anatomies and easier to recognize boundaries and appearance.

- *ExplicitClasses* consist of 28 classes including the blackground, surgical instruments and anatomical structures that occurred during the focued operation period. Due to its fine granularity, the class distribution is extremely skewed with 11 classes having less than 1% pixels among the entire datasets.

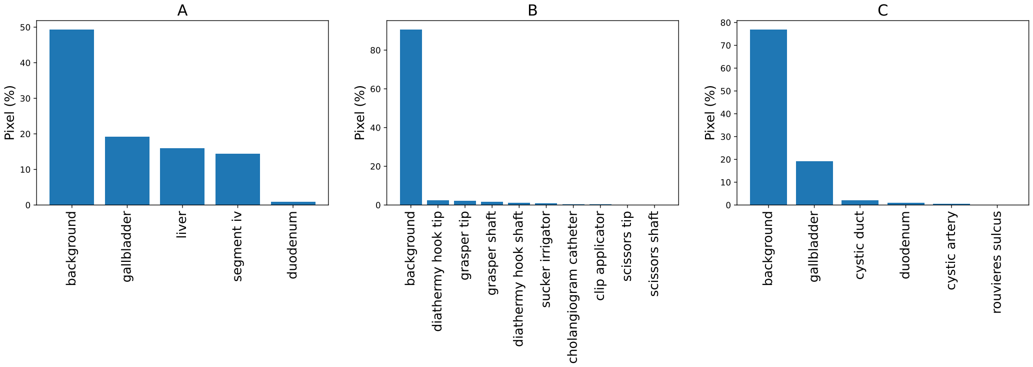


Figure 6: Class distribution of In-house Seg datasets with three different class granularity.

Table 5: Comparison of pre-training strategies on In-house Seg with different class granularity. *IoU* is reported in percentage. The best results are in **bold**.

Class definition	Class numbers	Performance
Only Instruments	10	67.46
Only Critical Anatomies	6	65.07
Only Easy Anatomies	5	71.82
Explicit Classes	28	44.80
Major Classes	20	62.26

From Table 5, we can observe that our method performs the best on Only easy anatomies, and the worst on Explicit classes. Intuitively, with more

Appendix D. Visualisation on public datasets

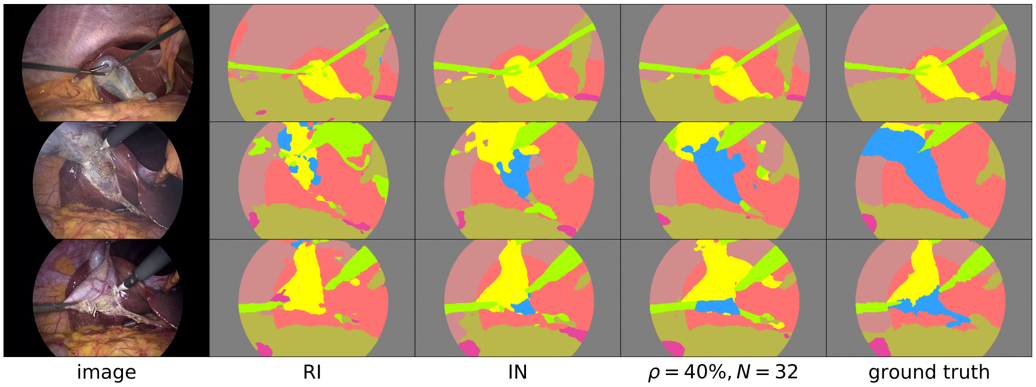


Figure 7: Predictions on three sample images from CholecSeg8k. From left to right shows the original images, predictions from no pre-training, supervised pre-training, our method, and the ground truth segmentation masks. The colour code follows the original dataset.

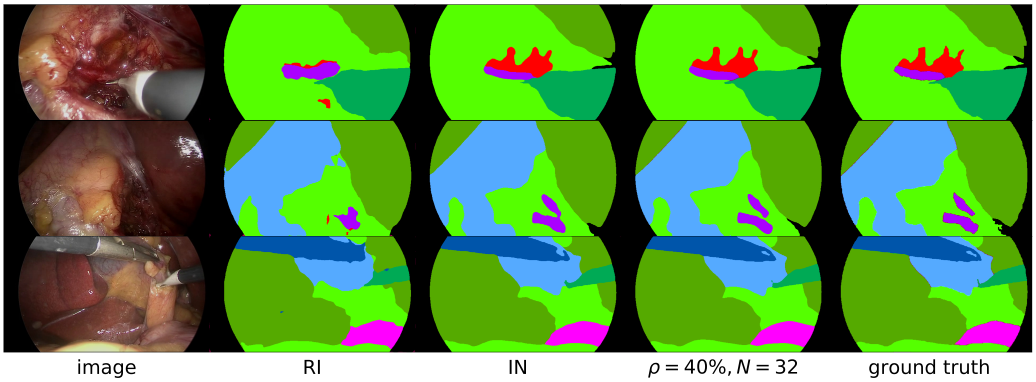


Figure 8: Predictions on three sample images from m2caiSeg. From left to right shows the original images, predictions from no pre-training, supervised pre-training, our method, and the ground truth segmentation masks. The colour code follows the original dataset.