Intraoperative Segmentation through Deep Learning and Mask Post-processing in Laparoscopic Liver Surgery

James Borgars, Jibran Raja, Abhinav Ramakrishnan, Abdul Karim Abbas, Aodhan Gallagher, Ahmad Najmi Mohamad Shahir, Theodora Vraimakis, and Sharib Ali*

School of Computer Science, University of Leeds, Leeds, United Kingdom
`s.s.ali@leeds.ac.uk`

Abstract. Laparoscopic liver surgery is a popular surgical approach due to its capabilities of minimising trauma, complications, and recovery times. The use of a laparoscope allows for developments in the field of machine-assisted surgery due to the availability of intraoperative imagery. Accurate landmark detection of the liver using laparoscopic footage is a dependency to many developments, such as 3D-2D registration. In this paper, we present experimental results measuring the suitability of popular segmentation models, and their compatibility with different loss functions when handling intraoperative images; we also present a pipeline in training models for this segmentation task, including a novel step of applying post-processing techniques to maximise accuracy. Our results are evaluated using precision, Dice similarity coefficient, and a symmetric distance metric. Our results show that through the use of our proposed pipeline, models retain their ability to generalise, and can lead to noticeably improved accuracy both quantitatively and qualitatively. We demonstrate the feasibility of utilising post-processing to improve predictions. Finally, possible future directions in this field following from our results are discussed. The code from this research has been made available and can be accessed here: https://github.com/ARMADILLO-VISION/SLiPPA

Keywords: Liver laparoscopy · Image segmentation · Deep learning · Post-processing

1 Introduction

1.1 Overview

Laparoscopic liver surgery, also known as minimally-invasive liver surgery, has emerged as a popular approach due to its reduction of patient trauma, recovery times, and complications compared to other approaches [18]. To enhance surgical precision in these procedures, machine-assisted approaches have long been an

* Corresponding author

area of laparoscopy research. An example of this is 3D-2D registration [10], where a 3D preoperative liver model is deformed, and key surgical landmarks, such as tumours and vessels, are projected onto intraoperative footage from a laparoscope [11]. A key component of the 3D-2D registration pipeline is that of 2D segmentation, where landmarks such as the falciform ligament and ridge [10], are detected in real-time during the surgery such that the current form of the liver is recorded. Prior research has demonstrated the efficacy of different model architectures and approaches to data augmentation with varying levels of success [2,10], however an approach achieving an acceptable level of accuracy is yet to be found [2,13]. In this paper we present ablation results into the suitability of five popular image segmentation models with prior use in surgical research [2,7,9], and propose a pipeline to train a model for liver segmentation (see Figure 1), utilising pre-trained weights, training, fine-tuning, followed by a novel mask post-processing step, thereby facilitating improved model performance.

1.2 Challenges

Obtaining consistently accurate predictions from deep learning models is a non-trivial task due to a number of reasons. Firstly, the liver is prone to deforming based on its environment and forces applied upon it [2]. This means that the liver shape can change drastically not only from patient to patient, but also the same liver through the duration of the surgical procedure. This issue, combined with the small amounts of available annotated data and a complex visual environments, leads to the necessity of dataset augmentations when training models [10]. As this task identifies linear landmarks, this leads to heavy class imbalance (see Table 1), with over 98% being labelled as background within the P2ILF training set. To counter this imbalance, the choice of loss function and class weights in this work must be selected appropriately.

Table 1. Occurrences of each class within the P2ILF training set.

Class	Count	% (3 s.f.)
Background	244,679,040	98.3
Silhouette	1,872,457	0.752
Ridge	1,765,372	0.709
Ligament	515,071	0.207

2 Related Work

2.1 Medical Image Segmentation

The introduction of UNet by Ronneberger et al. allowed for developments within the field by proposing the "U-shaped architecture" for Fully Convolutional Networks [17], in which a model has a contracting path with pooling layers and an expansive path with up-convolutions, with these paths connected by a bottleneck and skip connections. Models building upon the UNet architecture include UNet++ by Zhou et al. [22], a denser architecture with a greater number of convolution blocks and dense skip connection pathways, as well as the use of deep supervision; UNet3+ builds upon the the architecture of UNet++ [7], with its proposed use being within medical image segmentation. UNet3+ proposes full-scale skip connections where each convolution block in the contracting path has skip connections to its equivalent and below blocks in the expansive path; the bottleneck and each block in the expansive path is supervised by the ground truth, as well as having skip connections to every block further along the expansive path. The aforementioned models have all been measured against the LiTS 2017 benchmark [4,7], highlighting the focus on the medical imaging field. ResUNet is a deep residual UNet model which replaces the standard Convolution-ReLU block with a residual block with batch normalisation [21]. ResUNet++ builds on top of ResUNet, adding squeeze-excitation blocks for dynamic weighting of convolutional channels, ASPP to allow for broader context when classifying a pixel, and attention to enhance feature quality [8]. Tailored for medical image segmentation, ResUNet++ outperforms both UNet and ResUNet in colonoscopy segmentation benchmarks [8].

2.2 Laparoscopic Segmentation

Anteby et al. discuss the suitability of deep learning, notably Convolutional Neural Networks (CNNs), in the segmentation of laparoscopic imagery, having already revolutionised the field of medical imagery [3]. Applications such as tool detection and anatomy recognition were found to be suitable [3], with use cases only increasing as developments are made within the field. Koo et al. successfully demonstrated semantic contour detection of the ridge and silhouette of the liver through the use of CNNs, using CASENet with a ResNet50 encoder pre-trained on the ImageNet dataset [10]. Dataset augmentation through scale, shear, brightness, contrast, rotation, and translation were applied, promoting generalisation and invariance of the model with a small dataset [10]. As part of MICCAI 2022, the Preoperative to Intraoperative Laparoscopy Fusion (P2ILF) challenge was hosted, focusing on solving the end-to-end task of 3D-2D liver registration without human annotation, including liver landmark segmentation from laparoscopic images [2]. As opposed to the work presented by Koo et al., the P2ILF dataset also contained annotations for the falciform ligament [2,10]. Teams from around the world competed in the P2ILF challenge, covering a range of different approaches in terms of model, loss function, data augmentation, and pre-training [2]. Pei et al. introduce D^2GPLAND, a depth-aware model which is guided by unified features from an estimated depth map through the use of a depth estimation network and a Segment Anything Model (SAM) encoder, as well as using the ResNet34 encoder on the original image [15], achieving best-in-class results evaluating against their L3D dataset. Pei et al. publicly released L3D, which is the collation of laparoscopic images from multiple sources [15].

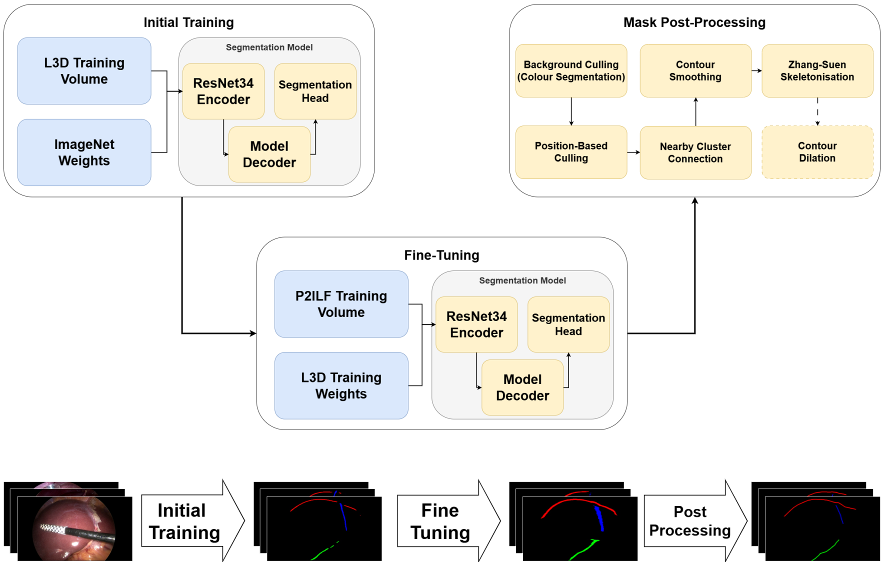

Fig. 1. Pipeline of 2D segmentation task. Dashes represent an optional process.

3 Methodology

3.1 Proposed Method

Firstly, all models selected for the ablation study – UNet, UNet++, UNet3+, DeepLabV3+, and ResUNet++ – are all equipped with ResNet34 encoders that have been pre-trained on the ImageNet dataset prior to training. The ResNet family of encoders have achieved state-of-the-art Dice scores on endoscopic segmentation [6], demonstrating its suitability. ResNet34 was chosen due to its balance of accuracy when compared to larger ResNet models [14], with less computational cost. Parameters ablated over include learning rate, batch size, and loss function. The model is trained on the L3D dataset, which has been augmented with flip transformations on both axes. The AdamW optimiser is used for training, and the loss functions ablated consist of different weightings of cross-entropy loss, Dice loss, Huber loss, and Focal-Tversky loss (see Equation 2). For cross-entropy loss, the following novel modified logarithmic class weights function was used, where M is the set of all training mask pixels, and $M_i \subseteq M$ is the set of all training mask pixels of class i:

$$W_i = \max\left[1.0, \log\left(\frac{|M|}{|M_i|}\right)\right] \tag{1}$$

Once training has been completed, models were evaluated using two validation patients chosen from the P2ILF training set, using precision and Dice similarity coefficient (DSC) metrics. Candidate models are then selected to be fine-tuned on the P2ILF dataset from this evaluation.

Finally, predicted masks from the fine-tuned model are post-processed. This involved culling known false positive pixels (i.e. pixels out of camera, ligament predictions above/below both the ridge and silhouette), applying the Ramer-Douglas-Peucker algorithm [16] for smoothing landmark contours, and applying the Zhang-Suen algorithm for skeletonisation of contours [20]. This centre line can then be dilated for evaluation, or sampled into a set of points. The Ramer-Douglas-Peucker algorithm is as follows:

1. The first and last points of the curve are connected with a straight line.
2. Every other point in the curve has its perpendicular distance to this straight line calculated.
3. If the furthest distance recorded is greater than some value ϵ, then the point this was found at is preserved, and two sub-curves are defined either side of this point.
4. The algorithm is recursively called for the two sub-curves that were previously defined.

3.2 Loss Function

Our proposed composite loss function using weighted cross-entropy ($\mathcal{L}_{\mathrm{wCE}}$), Dice loss ($\mathcal{L}_{\mathrm{D}}$), Huber loss ($\mathcal{L}_{\mathrm{H}}$), and Focal-Tversky loss ($\mathcal{L}_{\mathrm{FTL}}$) can be represented as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\mathrm{wCE}} + \beta \cdot \mathcal{L}_{\mathrm{D}} + \gamma \cdot \mathcal{L}_{\mathrm{H}} + \delta \cdot \mathcal{L}_{\mathrm{FTL}} \tag{2}$$

Here, α, β, γ, δ are the weights (i.e. hyperparameters) which we intend to ablate to find the optimal loss function.

4 Results

4.1 Dataset

Two datasets were used for model training: the L3D dataset was used to initially train the model [15], and the P2ILF dataset was used for further fine-tuning [2]. Three landmarks are present within the segmentation masks of both datasets: the silhouette, the ridge, and the falciform ligament.

The L3D dataset contains 1,152 image frames from 39 patients (122 frames are in the validation set, and a further 109 images are within the test set). The P2ILF dataset has 197 frames from 11 patients (47 images from two patients make up a validation set, and 30 images consisting of two patients make up a holdout set).

4.2 Evaluation Metrics

To evaluate model performance, the metrics used in the P2ILF challenge 2D segmentation task [2] were also used for our evaluation, these metrics being:

- Precision (P): this metric focuses on penalising false positives within the dataset to ensure that correct predictions are made.
- Dice similarity coefficient (D): this metric is used to focus on the similarity between the predicted segmentation mask (Y_{pred}) and the ground truth segmentation mask (Y_{truth}). It is the number of true positive predicted pixels multiplied by two and divided by the sum of predicted positive pixels and actual positive pixels.
$$D = \frac{2 \cdot |Y_{\text{pred}} \cap Y_{\text{truth}}|}{|Y_{\text{pred}}| + |Y_{\text{truth}}|} \tag{3}$$
- Symmetric distance (G): Ali et al. [2] use the symmetric distance proposed by François et al. [5]. $d_{\max}$ is a threshold value for whether a predicted landmark is spurious, B_I is the set of predicted image landmarks, whilst C_I is the set of ground truth image landmarks. Q is the tolerance region around the ground truth landmarks (defined by the threshold $d_{\max}$), and d_S is a function that calculates symmetric distance.

$$G = \frac{1}{2 \cdot |C_I| \cdot d_{\max}} \left(\sum_{b_I \in B_I \cap Q} d_S(b_I, C_I \backslash \text{FN}) + \sum_{c_I \in C_I \backslash \text{FN}} d_S(c_I, B_I \cap Q) \right) \tag{4}$$
$$+ \frac{|\text{FP}|}{|I| - 2 \cdot |C_I| \cdot d_{\max}} + \frac{|\text{FN}|}{|C_I|}$$

4.3 Experimental Setup

In training, the L3D dataset was augmented with mirror augmentations across both the $x-$ and $y-$ axes, followed by resizing to 416×320 pixels. The AdamW optimiser was used with a learning rate plateau scheduler, which multiplied learning rate by 0.2 after three epochs of stagnation. Training was performed on an NVIDIA RTX 4070, except for UNet3+, where an NVIDIA L40S was used due to extra VRAM being required. For the ablation study, learning rates of $\{0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ were tested, and batch sizes of $\{2, 4, 8, 16, 32\}$ were used (UNet3+ was only ablated up to batch size 8 due to VRAM constraints). A patience of 7 was used throughout.

For loss function experimentation, 35 ablations were conducted per model on the composite loss function in Equation 2. The loss function hyperparameters (α, β, γ, and δ) shown in Table 3 (see 5th column) are the best performing configurations from the ablation after other parameters were ablated over. Performance was evaluated using patients 1 and 2 from the training set of P2ILF as an evaluation set, providing us with candidate models for fine-tuning (see Table 3).

Table 2. P2ILF 2D segmentation challenge results [2] evaluated using our implementations of precision, Dice similarity coefficient, and the symmetric distance metric. Results are in the order of ridge, falciform ligament, and silhouette. $\bar{P}$, $\bar{D}$, $\bar{G}$ are the metric mean across all three landmarks. The best mean results are highlighted in bold.

Team	**P** ↑	$\bar{\mathbf{P}}$ ↑	**D** ↑	$\bar{\mathbf{D}}$ ↑	**G** ↓	$\bar{\mathbf{G}}$ ↓
BHL	0.24/0.41/0.46	**0.37**	0.22/0.43/0.50	**0.38**	0.70/0.43/0.40	0.51
NCT	0.20/0.31/0.41	0.31	0.24/0.32/0.52	0.36	0.52/0.51/0.32	**0.45**
UCL	0.11/0.43/0.38	0.31	0.13/0.48/0.40	0.34	0.73/0.63/0.42	0.59
VIP	0.11/0.23/0.19	0.18	0.16/0.33/0.29	0.26	0.71/0.44/0.62	0.59
VOR	0.10/0.15/0.16	0.13	0.15/0.24/0.25	0.21	0.70/0.65/0.66	0.67

Models were fine-tuned using the P2ILF training set (patients 1 and 2 were used as a validation set) with a learning rate of 0.0001, a batch size of 8, and a patience of 7. Two rounds of fine-tuning was performed: once with solely cross-entropy loss and another with a loss function consisted of weighted cross-entropy loss ($\mathcal{L}_{\text{wCE}}$) and Focal-Tversky loss ($\mathcal{L}_{\text{FTL}}$) with $\alpha = 0.75$ and $\delta = 0.25$ (see Equation 2) A Focal-Tversky component was added to the fine-tuning loss function due to low recall of models after fine-tuning with solely cross-entropy loss (shown in Figure 2). The Tversky index represents a generalisation of Dice loss, with the focal component promoting predictions of more difficult classes (i.e. landmarks) [1]. Lahlouh et al. achieve their best results using a combination of cross-entropy loss and Focal-Tversky loss when performing segmentation on cerebral angiography imagery [12], highlighting its medical suitability.

Table 3. Candidate models selected from the L3D training ablation study. Here, α, β, γ, and δ are the coefficients used in the loss function $\mathcal{L}$ in Equation 2.

No.	**Architecture**	**Learning Rate**	**Batch Size**	$\alpha/\beta/\gamma/\delta$
1	UNet	0.0001	32	1.00/0.00/0.00/0.00
2	UNet	0.0001	8	1.00/0.00/0.00/0.00
3	UNet++	0.001	32	0.25/0.25/0.00/0.50
4	UNet++	0.0005	16	0.50/0.00/0.50/0.00
5	UNet++	0.0005	16	0.75/0.25/0.00/0.00
6	UNet3+	0.001	8	0.50/0.00/0.00/0.50
7	UNet3+	0.001	8	0.50/0.25/0.00/0.25
8	DeepLabV3+	0.0001	64	0.25/0.00/0.25/0.50
9	ResUNet++	0.0005	16	0.50/0.00/0.00/0.50
10	ResUNet++	0.0005	8	0.75/0.00/0.00/0.25

4.4 Quantitative Results

Firstly, testing of the cross-entropy weighting function was performed, the same model with the same parameters was trained with three different class weightings for cross-entropy loss: proportional, logarithmic weighting, and our custom weighting (see Equation 1), fine-tuning of models and post-processing was not performed on any evaluations. As shown in Table 4, standard proportional weighting simply fails at the segmentation task, logarithmic weights provide a significant improvement, but by using our novel class weighting function, precision is doubled compared to logarithmic weights, and best performance in all three evaluation metrics.

Table 4. Evaluation of different cross-entropy class weightings on the P2ILF test set. Best results are shown in bold.

Weighting	$\bar{\mathbf{P}} \uparrow$	$\bar{\mathbf{D}} \uparrow$	$\bar{\mathbf{G}} \downarrow$
Proportional	0.05	0.00	1.00
Logarithmic	0.19	0.22	0.76
Ours	**0.41**	**0.29**	**0.54**

Table 3 shows the configurations of the ten best candidate models from the ablation study. For this, at least one model from each architecture was selected based on highest combined mean precision and DSC from the P2ILF evaluation set. Table 2 illustrates the P2ILF challenge results using our evaluation implementation for comparison.

Table 5 shows the performance of candidate models after initial training, fine-tuning, and post-processing, evaluating against the P2ILF test set. Fine-tuning was done with solely weighted cross-entropy loss ($\mathcal{L}_{\text{wCE}}$). 50% of models outperformed P2ILF in mean precision prior to any fine-tuning (i.e. solely trained on ImageNet followed by L3D, never having seen the P2ILF dataset), with this proportion increasing to 80% after fine-tuning. However, no model at any points outperforms the P2ILF teams in mean DSC or symmetric distance. Post-processing sees to have minimal effect on evaluation in Table 5, with slight increases in mean DSC and precision on average, but with worse performance in symmetric distance.

Table 6 shows the evaluation after the Focal-Tversky component was added to the loss function for fine-tuning. It was noted that this resulted in better performance compared to Table 5 with regards to the DSC and symmetric distance metrics after fine-tuning, but did not consistently outperform P2ILF in any metric (10% of models in mean precision and symmetric distance, and 40% of models in mean DSC) prior to post-processing.

Table 5. Results of candidate model fine-tuning using cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$). Highlighted cells represent a result that beats all P2ILF teams in that metric, the best results are highlighted in bold.

Candidate	Initial Training			Fine-Tuning			Post-Processing		
	$\bar{P}_{\texttt{init}} \uparrow$	$\bar{D}_{\texttt{init}} \uparrow$	$\bar{G}_{\texttt{init}} \downarrow$	$\bar{P}_{\texttt{tune}} \uparrow$	$\bar{D}_{\texttt{tune}} \uparrow$	$\bar{G}_{\texttt{tune}} \downarrow$	$\bar{P}_{\texttt{post}} \uparrow$	$\bar{D}_{\texttt{post}} \uparrow$	$\bar{G}_{\texttt{post}} \downarrow$
1	0.39	0.22	0.69	0.45	0.23	0.67	0.44	0.26	0.67
2	0.41	0.29	0.54	0.44	0.33	0.52	0.45	**0.35**	0.53
3	0.35	0.28	0.59	0.38	0.30	0.56	0.43	0.32	0.57
4	0.42	0.25	0.64	0.43	0.30	0.55	0.45	0.32	0.57
5	0.39	0.28	0.61	0.38	0.31	0.56	0.39	0.32	0.57
6	0.36	0.24	0.60	0.51	0.31	**0.51**	**0.52**	0.33	0.54
7	0.32	0.31	0.56	0.42	0.25	0.62	0.44	0.28	0.63
8	0.34	0.23	0.65	0.36	0.21	0.69	0.36	0.22	0.72
9	0.38	0.27	0.63	0.30	0.15	0.80	0.30	0.15	0.79
10	0.32	0.17	0.70	0.43	0.17	0.70	0.45	0.21	0.72

Table 6. Results of candidate model fine-tuning using a combined loss function of cross-entropy loss ($\mathcal{L}_{\mathrm{CE}}$) and Focal-Tversky loss ($\mathcal{L}_{\mathrm{FTL}}$). Highlighted cells represent a result that beats all P2ILF teams in that metric, the best results are highlighted in bold.

Candidate	Initial Training			Fine-Tuning			Post-Processing		
	$\bar{P}_{\texttt{init}} \uparrow$	$\bar{D}_{\texttt{init}} \uparrow$	$\bar{G}_{\texttt{init}} \downarrow$	$\bar{P}_{\texttt{tune}} \uparrow$	$\bar{D}_{\texttt{tune}} \uparrow$	$\bar{G}_{\texttt{tune}} \downarrow$	$\bar{P}_{\texttt{post}} \uparrow$	$\bar{D}_{\texttt{post}} \uparrow$	$\bar{G}_{\texttt{post}} \downarrow$
1	0.39	0.22	0.69	0.35	0.37	0.47	0.45	0.37	0.44
2	0.41	0.29	0.54	0.38	0.40	0.40	**0.48**	**0.43**	**0.39**
3	0.35	0.28	0.59	0.35	0.37	0.47	0.47	0.39	0.45
4	0.42	0.25	0.64	0.31	0.36	0.53	0.40	0.37	0.48
5	0.39	0.28	0.61	0.36	0.39	0.46	0.44	0.39	0.45
6	0.36	0.24	0.60	0.33	0.41	0.47	0.44	0.42	0.40
7	0.32	0.31	0.56	0.33	0.40	0.54	0.45	0.42	0.43
8	0.34	0.23	0.65	0.31	0.29	0.60	0.40	0.30	0.56
9	0.38	0.27	0.63	0.31	0.29	0.66	0.35	0.28	0.56
10	0.32	0.17	0.70	0.32	0.34	0.51	0.39	0.34	0.50

It is evident that post-processing led to noticeable improvements for this model. Mean precision increased on average over 9%, from 33.5% to 42.7%, making 90% of candidate models outperform all P2ILF teams in this metric. Improvements were also seen across DSC and symmetric distance. From this evaluation, there are five candidate models that match or beat all teams from the P2ILF 2D segmentation task in all metrics, with three models outperforming in every metric outright (see Table 2 and Table 6).

Candidate 2 performed best: the model provided an 11% increase in mean precision (30% relative increase), a 5% increase in mean Dice score (over 13% relative increase), and 6% decrease in symmetric distance (over 13% relative improvement) compared to the best result in each metric in the P2ILF 2D segmentation task (see Table 2 and Table 6).

Regarding execution performance, the time taken for inference and post-processing was recorded. Inference ranged from 9-15 milliseconds, whilst post-processing added a 9-12 millisecond penalty. This results in a production frequency range of 37-56Hz, acceptable for real-time operation.

4.5 Qualitative Results

Figure 2 shows six example images from the P2ILF test set, three from each test patient (patients 4 and 11). Results are shown from the two best performing teams reported in the P2ILF challenge 2D segmentation task (BHL and NCT - see Table 2) together with our best performing model - candidate 2. Resulting inferences from initial training of the model (see `TRAIN` in Figure 2), P2ILF fine-tuned models without post-processing (see `CE_FT` and `FTL_FT`), and finally with post-processing (see `CE_PP` and `FTL_PP`) are shown. Results that have the `CE_` prefix is where candidate 2 was fine-tuned solely with cross-entropy loss; results that have the `FTL_` prefix is where candidate 2 was fine-tuned with both a cross-entropy loss component and a Focal-Tversky loss component, as described in 4.3.

Our qualitative results demonstrate that the approaches employed by BHL and NCT can result in broken contours, particularly noticeable on images from patient 11. Sub-figure `TRAIN` shows that the model under-predicts landmarks, albeit with high precision when it does predict (i.e. a low false positive rate). `CE_FT` and `PP_FT` still exhibit this behaviour, but this does not seem to be present in `FTL_FT`. This increase does lead to thicker contours however, with some contours being disconnected or broken. Once `FTL_FT` has been post-processed, shown in `FTL_PP`, contours are now connected and consistent in thickness, cleaning up the output of the model significantly.

Figure 3 shows two failure cases for the model, it is evident that the model has difficulty predicting in cases where there is a low level of lighting, and where the liver is being manipulated such that the anterior view is not fully facing the camera. The precision in these images are still of good quality, reinforcing the idea that the model is able to generalise well, but cannot predict the full contours in extreme cases, although maintaining a low false positive rate from a qualitative perspective.

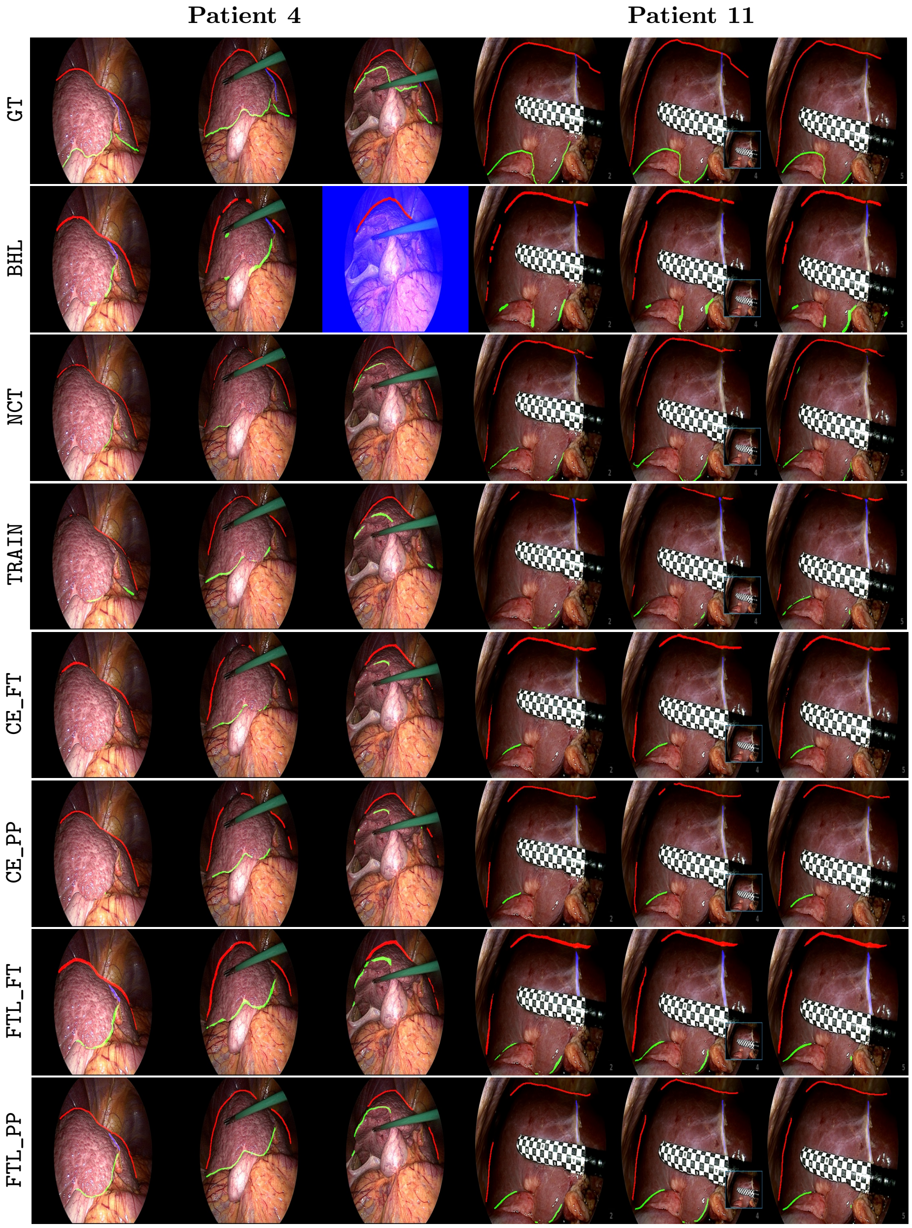

Fig. 2. Qualitative results on the P2ILF test set. `GT` shows ground truth annotations. `BHL` and `NCT` are the predictions made by the BHL and NCT teams from the P2ILF challenge respectively [2]. `TRAIN` is the predictions made by candidate 2 after initial training. `CE_FT` represents candidate model 2 being fine-tuned solely with cross-entropy loss (without post-processing). `CE_PP` is the same model as `CE_FT` but with post-processing applied. `FTL_FT` represents candidate model 2 being fine-tuned with both cross-entropy loss and Focal-Tversky loss (without post-processesing). `FTL_PP` is the same model as `FTL_CE` but with post-processing applied.

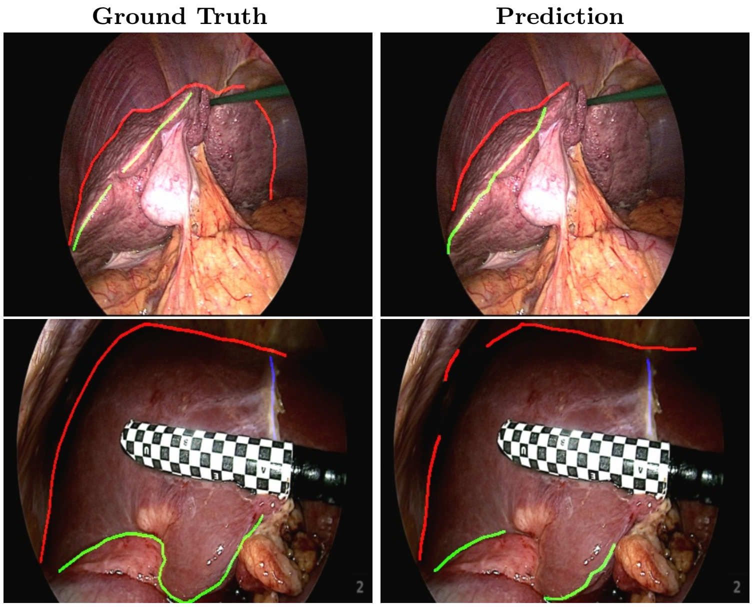

Fig. 3. Examples of failure cases of candidate 2, compared with ground truth masks.

5 Discussion & Conclusion

5.1 Discussion

Table 5 and `CE_FT` in Figure 2 show that the model is precise when making a prediction, however it is visible that the recall is low. Due to this under-prediction, the advantages of post-processing is minimised (see `CE_PP`) due to thin and sparse predictions meaning there is little to process, the difference between `FTL_FT` and `FTL_PP` is visibly greater. The lacklustre post-processing performance in `CE_PP` is also seen quantitatively, with `tune` and `post` values being noticeably similar in Table 5, when compared to differences shown in Table 6.

Table 6 and sub-figure `FTL_FT` in Figure 2 show a model that is less prone to under-prediction when compared to `CE_FT`, correctly predicting a greater proportion of landmarks, albeit with thicker contours. `FTL_PP` visualises the improvement in metrics shown in Table 6 after post-processing, with thinner contours, connection of nearby contours, whilst also leading to a more accurately predicted contour as shown by the improved DSC and symmetric distance scores.

Adding a Focal-Tversky loss component to the loss function promotes prediction of landmarks within a mask due to the focal exponent adding emphasis to landmark predictions that are of low certainty (i.e. difficult predictions). Post-processing improves evaluation metrics when there is over-prediction present by thinning contours such that they are more precise, with smoothing simplifying the contour such that the jaggedness of a contour is removed, due to it not reflecting the form shown by the landmarks of the liver that the model is attempting to predict.

The approach taken by Pei et al. in the estimation of depth masks attempts to provide more context to the model for predictions [15], an approach that we believe can be studied further. Labrunie et al. utilise distance maps as opposed to binary masks when training their 3D-2D registration model [11], allowing models to learn data that does not have harsh binary boundaries around the thin contours, and is highly transferable to the task of intraoperative segmentation of the liver through these landmarks. The mask post-processing presented in this paper lays out the fundamentals of this approach, with in-depth ablation studies, and ideas such as mask correction through deep learning yet to be explored.

There is a question concerning the appropriateness of DSC as an evaluation metric, due to the task predicting contours, rather than segments of an image; alternatives such as centre line Dice (clDice) [19] should have their suitability investigated in this field.

5.2 Conclusion

In this work, we proposed a systematic approach in training segmentation models in the context of liver laparoscopy. The methodology involves utilising pretrained models on a large generic dataset, such as ImageNet, followed by further training on a large dataset of laparoscopic images, which can be obtained from multiple sources as demonstrated by the L3D dataset. Fine-tuning can then be performed on a high quality dataset from a single source. Furthermore, we introduce a novel post-processing pipeline that incorporates colour segmentation, skeletonisation, and contour smoothing to minimise errors from false positive predictions. We present results that outperform in all metrics used for the P2ILF challenge 2D segmentation task, on the same dataset. Research areas such as synthetic data generation and depth estimation hold promise for development in this field. The research presented in this work forms part of a 3D-2D registration pipeline, where no human annotation is required.

Acknowledgements. This project is part of the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC) pump-priming funding as the "ARMADILLO" project (NIHR203331). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abraham, N., Khan, N.M.: A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). pp. 683–687 (2019). https://doi.org/10.1109/ISBI.2019.8759329
2. Ali, S., Espinel, Y., Jin, Y., Liu, P., Güttner, B., Zhang, X., et al.: An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion from the MICCAI2022 challenge. Medical Image Analysis **99**, 103371 (2025). https://doi.org/10.1016/j.media.2024.103371
3. Anteby, R., Horesh, N., Soffer, S., Zager, Y., Barash, Y., Amiel, I., et al.: Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. Surgical Endoscopy **35**(4), 1521–1533 (Apr 2021). https://doi.org/10.1007/s00464-020-08168-1
4. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023). https://doi.org/10.1016/j.media.2022.102680
5. François, T., Calvet, L., Zadeh, S.M., Saboul, D., Gasparini, S., Samarakoon, P., et al.: Detecting the occluding contours of the uterus to automatise augmented laparoscopy: score, loss, dataset, evaluation and user study. International Journal of Computer Assisted Radiology and Surgery **15**, 1177–1186 (2020). https://doi.org/10.1007/s11548-020-02151-w
6. Ghamsarian, N., Wolf, S., Zinkernagel, M., Schoeffmann, K., Sznitman, R.: DeepPyramid+: medical image segmentation using pyramid view fusion and deformable pyramid reception. International Journal of Computer Assisted Radiology and Surgery **19**(5), 851–859 (May 2024). https://doi.org/10.1007/s11548-023-03046-2
7. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al.: UNet 3+: A full-scale connected UNet for medical image segmentation. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1055–1059 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053405
8. Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., Lange, T.D., Halvorsen, P., et al.: ResUNet++: An advanced architecture for medical image segmentation. In: 2019 IEEE International Symposium on Multimedia (ISM). pp. 225–2255 (2019). https://doi.org/10.1109/ISM46123.2019.00049
9. Kojima, S., Kitaguchi, D., Igaki, T., Nakajima, K., Ishikawa, Y., Harai, Y., et al.: Deep-learning-based semantic segmentation of autonomic nerves from laparoscopic images of colorectal surgery: an experimental pilot study. International Journal of Surgery **109**(4) (2023). https://doi.org/10.1097/JS9.0000000000000317
10. Koo, B., Robu, M.R., Allam, M., Pfeiffer, M., Thompson, S., Gurusamy, K., et al.: Automatic, global registration in laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. **17**(1), 167–176 (2022). https://doi.org/10.1007/s11548-021-02518-7
11. Labrunie, M., Ribeiro, M., Mourthadhoi, F., Tilmant, C., Le Roy, B., Buc, E., et al.: Automatic preoperative 3D model registration in laparoscopic liver resection. Int. J. Comput. Assist. Radiol. Surg. **17**, 1429–1436 (2022). https://doi.org/10.1007/s11548-022-02641-z
12. Lahlouh, M., Blanc, R., Piotin, M., Szewczyk, J., Passat, N., Chenoune, Y.: Cerebral AVM segmentation from 3D rotational angiography images by convolutional neural networks. Neuroscience Informatics **3**(3), 100138 (2023). https://doi.org/10.1016/j.neuri.2023.100138

13. Lin, Y., Xu, J., Hong, J., Si, Y., He, Y., Zhang, J.: Prognostic impact of surgical margin in hepatectomy on patients with hepatocellular carcinoma: A meta-analysis of observational studies. Front. Surg. **9**, 810479 (Feb 2022). https://doi.org/10.3389/fsurg.2022.810479
14. Pamungkas, Y., Triandini, E., Yunanto, W., Thwe, Y.: Impact of hyperparameter tuning on ResNet-UNet models for enhanced brain tumor segmentation in MRI scans. International Journal of Robotics and Control Systems **5**(2), 917–936 (2025). https://doi.org/10.31763/ijrcs.v5i2.1802
15. Pei, J., Cui, R., Li, Y., Si, W., Qin, J., Heng, P.A.: Depth-Driven Geometric Prompt Learning for Laparoscopic Liver Landmark Detection . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15006. Springer Nature Switzerland (October 2024). https://doi.org/10.1007/978-3-031-72089-5$_1$5
16. Ramer, U.: An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing **1**(3), 244–256 (1972). https://doi.org/10.1016/S0146-664X(72)80017-0
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing (2015). https://doi.org/10.1007/978-3-319-24574-4$_2$8
18. Schneider, C., Allam, M., Stoyanov, D., Hawkes, D.J., Gurusamy, K., Davidson, B.R.: Performance of image guided navigation in laparoscopic liver surgery - a systematic review. Surg. Oncol. **38**(101637), 101637 (2021). https://doi.org/10.1016/j.suronc.2021.101637
19. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., et al.: clDice - a novel topology-preserving loss function for tubular structure segmentation. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16555–16564 (2021). https://doi.org/10.1109/CVPR46437.2021.01629
20. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Commun. ACM **27**(3), 236–239 (Mar 1984). https://doi.org/10.1145/357994.358023
21. Zhang, Z., Liu, Q., Wang, Y.: Road extraction by deep residual U-Net. IEEE Geoscience and Remote Sensing Letters **15**(5), 749–753 (2018). https://doi.org/10.1109/LGRS.2018.2802944
22. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: A nested U-Net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. p. 3–11. Springer-Verlag (2018). https://doi.org/10.1007/978-3-030-00889-5_1