

Joshua Sleeman*, Lorena Krames, and Werner Nahm

Towards Liver Segmentation in Laparoscopic Images by Training U-Net With Synthetic Data

<https://doi.org/10.1515/cdbme-2024-2147>

Abstract: The lack of labeled, intraoperative patient data in medical scenarios poses a relevant challenge for machine learning applications. Given the apparent power of machine learning, this study examines how synthetically-generated data can help to reduce the amount of clinical data needed for robust liver surface segmentation in laparoscopic images. Here, we report the results of three experiments, using 525 annotated clinical images from 5 patients alongside 20,000 synthetic photo-realistic images from 10 patient models. The effectiveness of the use of synthetic data is compared to the use of data augmentation, a traditional performance-enhancing technique. For training, a supervised approach employing the U-Net architecture was chosen. The results of these experiments show a progressive increase in accuracy. Our base experiment on clinical data yielded an F_1 score of 0.72. Applying data augmentation to this model increased the F_1 score to 0.76. Our model pre-trained on synthetic data and fine-tuned with augmented data achieved an F_1 score of 0.80, a 4% increase. Additionally, a model evaluation involving k-fold cross validation highlighted the dependency of the result on the test set. These results demonstrate that leveraging synthetic data has the ability of limiting the need for more patient data to increase the segmentation performance.

Keywords: Artificial neural networks, U-Net, liver segmentation, synthetic data

1 Introduction

Minimally invasive interventions offer numerous advantages over open surgery [1]. However, laparoscopic procedures on complex organs like the liver are challenging [2]. In recent years it has become apparent that machine learning has the power to aid the surgeon during operation, for example through semantic image segmentation [3]. Supervised training of an artificial neural network requires a large amount of labeled data with a diverse set of examples to ensure generaliza-

tion to a wide range of scenarios [4]. However, such datasets are often not available, especially in the medical domain. We investigate the use of synthetic data with the objective to alleviate the problem of insufficient clinical data [5]. Here, we report on ways of utilizing synthetic data by comparing its performance to models trained solely with clinical data. The training approach employs a supervised training using the U-Net architecture [6].

2 Methods

2.1 Datasets

We utilized two datasets in this study. The first, clinical dataset consists of 525 annotated images obtained from five patient videos and was provided by the Universitätsklinikum Köln. Each frame of the patient videos was annotated by hand, meaning a segmentation map containing segmentations of the liver and accordingly of the background was created for each frame. The second, synthetic dataset encompasses 20,000 images sourced from 10 patient models [5]. The synthetic images were generated to simulate photo-realistic laparoscopic images. To achieve this, images with a basic outline of a laparoscopic liver surgery scene were produced from a 3D CT liver dataset, which were used to train a setup of generative adversarial networks. Corresponding segmentation masks for the liver and background are supplied. For further information we refer to [5].

2.2 Approach

The approach involves supervised training of the U-Net architecture [6]. Because the clinical and synthetic data are structured into patient videos, it is the natural choice to choose the training, validation and test split by patient videos. Accordingly, one patient video was designated for the test set, another for the validation set and the remaining patient videos were allocated to the training set.

For the loss function, cross-entropy [7] was selected while the optimization algorithm used was adam [8], initialized with the default learning rate of $1 \cdot 10^{-3}$. The number of epochs for training was set using early-stopping [4].

We employed data augmentation to artificially increase the size of the clinical dataset, utilizing a combination of geometry and color space transformation techniques [4, 9]. Flipping and

*Corresponding author: Joshua Sleeman, Institute of Biomedical Engineering (IBT), Karlsruhe Institute of Technology (KIT), Fritz-Haber-Weg 1, 76131 Karlsruhe, Germany, e-mail: publications@ibt.kit.edu

Lorena Krames, Werner Nahm, IBT, KIT, 76131 Karlsruhe, Germany

random elastic deformation were selected as geometry transformation techniques. Random elastic deformation is considered as being the data augmentation technique which produces the best performance improvement [6]. For the clinical images, we determined the suitable values of the random elastic deformation to be 25 for sigma and the dimension of the kernel as 6×6 for the Gaussian filter [10]. Additionally, brightness changes were implemented. The data augmentation techniques were applied in a random fashion: On half of the images we used a combination of horizontal and/or vertical flipping and brightness increase or decrease by a random factor (maximum by half). On the other half we carried out random elastic deformation. The original images were retained in the training set, effectively doubling its size. All augmentation techniques were conducted prior to training, ensuring the random application of the techniques did not distort the results.

We implemented a 5-fold cross validation method for the clinical data, with each patient video serving as a fold. No validation set was used, ensuring the same conditions for all models. This methodology created five models, from which we derived the average results, avoiding an influence by the choice of the test set [4].

2.3 Experiments

The experiments can be categorized in three main parts: clinical, synthetic and combined experiments. First, the U-Net was trained just with the clinical data (**Exp. 1a**). Next, a new model was trained with augmented clinical data (**Exp. 1b**). Both models were tested on clinical data. In the second part, the U-Net was trained with synthetic data. This model was tested on both synthetic (**Exp. 2a**) and clinical (**Exp. 2b**) data. Lastly, a model pre-trained with synthetic data was fine-tuned with augmented clinical data (**Exp. 3a**). A 5-fold cross validation (**Exp. 3b**) was then performed as the final experiment. All models were tested on clinical data. Apart from the 5-fold cross validation experiment, all models were tested on the same patient video (clinical or synthetic) to ensure comparability between the models. The accuracy metrics used were recall (R), precision (P) and the F_1 score (F_1).

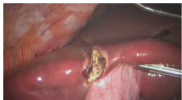


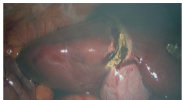


input image	ground truth	prediction	
			R 0.88
			P 0.93
			F_1 0.91
			R 0.51
			P 0.68
			F_1 0.58

Fig. 1: Experiment 1a. Model trained with clinical data and tested on clinical data.

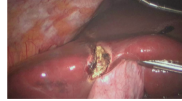


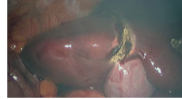

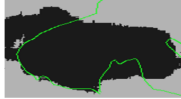
input image	ground truth	prediction	
			R 0.81
			P 0.86
			F_1 0.83
			R 0.68
			P 0.75
			F_1 0.71

Fig. 2: Experiment 1b. Model trained with augmented clinical data and tested on clinical data.







input image	ground truth	prediction	
			R 0.98
			P 0.98
			F_1 0.98
			R 0.99
			P 1.00
			F_1 0.99

Fig. 3: Experiment 2a. Model trained with synthetic data and tested on synthetic data.

3 Results

Figures 1 – 6 show a qualitative image analysis, where R , P and F_1 are given for each test image. The accuracy metrics stated in this section are the average values taken across the accuracy metrics of all images in the test set.

3.1 Clinical Experiments

Exp. 1a: Base Experiment The results for the qualitative image analysis are depicted in Figure 1. The prediction for the top image achieved values of 0.88, 0.93 and 0.91 for R , P and F_1 , respectively, while the prediction for the bottom image returned values of 0.51, 0.68 and 0.58. The average R , P and F_1 scores across the whole dataset were 0.67, 0.77 and 0.72, respectively (see Table 1a).

Exp. 1b: Data Augmentation In comparison, Figure 2 illustrates the qualitative image analysis for a model trained with data augmentation. For the top image, R , P and F_1 were 0.81, 0.86, and 0.83. Conversely, for the bottom image, these metrics were 0.68, 0.75, and 0.71. The average results across the entire test set (Table 1a) led to values of 0.74, 0.78, and 0.76 for R , P and F_1 , respectively.

3.2 Synthetic Experiments

Exp. 2a: Testing on Synthetic Data Figure 3 displays the qualitative image analysis. The prediction for the top image achieved a value of 0.98 for R , P and F_1 , while the prediction for the bottom image returned values of 0.99, 1.00 and 0.99. Table 1a presents the average accuracy metrics across

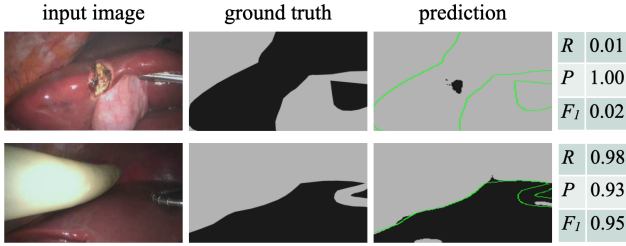


Fig. 4: Experiment 2b. Model trained with synthetic data and tested on clinical data.

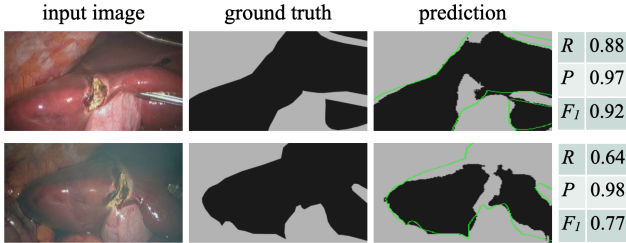


Fig. 5: Experiment 3a. Model pre-trained with synthetic data and fine-tuned with augmented clinical data. Tested on clinical data.

the dataset, indicating values of 0.96, 0.97, and 0.97 for R , P , and F_1 , respectively.

Exp. 2b: Testing on Clinical Data The results are illustrated in Figure 4. For the prediction of the top image, R , P and F_1 were 0.01, 1.00, and 0.02, respectively. Conversely, the prediction of the bottom image achieved values of 0.98, 0.93, and 0.95, respectively. Upon averaging the test results across the entire clinical dataset, the resulting values for R , P and F_1 were 0.12, 0.51, and 0.17, respectively (see Table 1a).

3.3 Combined Experiment

Exp. 3a: Pre-Training and Fine-Tuning Figure 5 depicts the results. The top image attained prediction results of 0.88, 0.97 and 0.92 for R , P and F_1 , respectively, while the bottom image reached results of 0.64, 0.98 and 0.77. The average accuracy metrics (Table 1a) were 0.69, 0.95 and 0.80.

Exp. 3b: 5-Fold Cross Validation The results for the 5-fold cross validation are illustrated in Figure 6. Table 1b gives the average accuracy metrics for each model tested with the specific test set. The values for the F_1 score ranged from 0.72 to 0.94, where patient 2 achieved the lowest and patient 4 the highest F_1 score. Upon averaging the test results of all 5 models, R reached a value of 0.83, P a value of 0.91 and the overall F_1 score was 0.85.

4 Discussion

The results garnered through our experiments highlight the capabilities of deep learning, more specifically the U-Net, to seg-

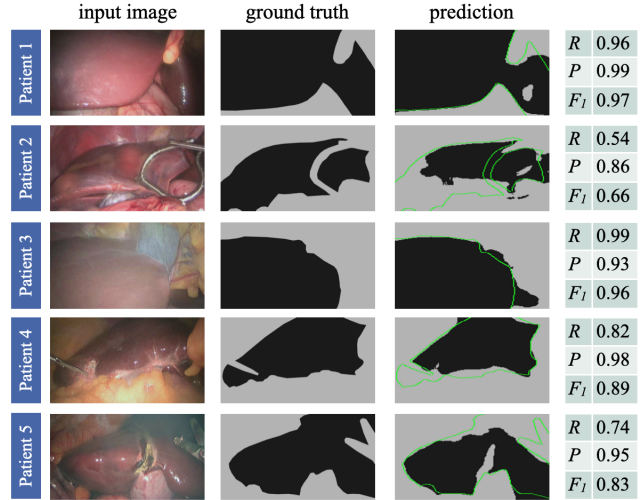


Fig. 6: Experiment 3b. 5-fold cross validation. Models pre-trained with synthetic data and fine-tuned with augmented clinical data. Tested on clinical data.

Tab. 1: Accuracy metrics of all experiments. The R , P and F_1 scores are given for each model. The result is the average value across the whole test dataset.

(a) Exp. 1a - Exp. 3a			
experiment	R	P	F_1 score
Exp. 1a	0.67	0.77	0.72
Exp. 1b	0.74	0.78	0.76
Exp. 2a	0.96	0.97	0.97
Exp. 2b	0.12	0.51	0.17
Exp. 3a	0.69	0.95	0.80
(b) Exp. 3b: 5-fold cross validation			
test patient	R	P	F_1 score
patient 1	0.85	0.94	0.88
patient 2	0.67	0.86	0.72
patient 3	0.98	0.87	0.92
patient 4	0.91	0.97	0.94
patient 5	0.75	0.90	0.81
average	0.83	0.91	0.85

ment the liver in laparoscopic images. This accomplishment could be achieved even with a small dataset.

The clinical experiment exemplifies the impact of using data augmentation for training. Despite notable success, challenges arose from reflections on the liver surface. Furthermore, incisions on the liver surface posed difficulties. This suggests the model struggles with unseen features, because the patient video used as a test set here is the only video containing an example of such an incision. Noteworthy here is also the absence of incisions in the synthetic dataset.

In contrast, the second test image exhibited a lower F_1 score, with challenges including misclassification of background as

liver due to similar appearances and complexities in darker regions. Data augmentation was explored in **Exp. 1b**, leading to varying results. While the first test image in Figure 2 experienced an 8% decrease in F_1 score, the second test image showed a 13% increase. Overall, a 4% improvement in F_1 score compared to without data augmentation could be reached. These findings suggest that while data augmentation can improve model robustness and accuracy, its effectiveness may vary across different scenarios. Striking the right balance between producing a variation of an image and realism is key here.

In the synthetic experiment, the effectiveness of training with synthetic data and its relevance to clinical scenarios is explored. The model showcased near-perfect segmentation when tested on synthetic images. This excellence in segmentation is seen across the whole synthetic test set, which achieved high average metric values. However, when tested on clinical data, the model did not perform well. Notably, some clinical images showed promising results, as exemplified by the second test image. This contrast in performance could potentially be attributed to the presence of a white ultrasonic head in the clinical image, which provides a strong contrast between the liver and the background. Nonetheless, these instances of remarkable predictions remained isolated cases. Thus, the results suggest a substantial domain gap between synthetic and clinical data, posing a significant challenge for accurate segmentation. This divergence becomes apparent upon visual comparison between the synthetic and clinical data. Despite the synthetic images having a similar appearance to the clinical images, key differences can be detected in liver color, surgical tool shape as well as the absence of incisions in the synthetic dataset.

A possible solution to this domain gap problem is the use of a pre-training approach, as conducted in **Exp. 3a**. The performance of a model pre-trained with synthetic data and fine-tuned with clinical data was compared to models trained solely with clinical data. Interestingly, the reflections on the liver surface were accurately classified, an accomplishment possibly attributable to the U-Net's exposure to a greater number of examples featuring reflections during the pre-training phase. However, accurately classifying the incisions on the liver remained an issue. Despite these challenges, a substantial improvement was observed in the average value compared to the results of **Exp. 1a**.

In the 5-fold cross-validation experiment, the model's performance was assessed using test images from various patient videos. Notable variations in F_1 scores were observed across different test patients. While some images achieved high accuracy with scores close to one, others exhibited lower scores due to challenges such as shadows and irregular surgical tools (e.g., patient 2 and 5). Interestingly, simpler layouts and clear liver-background distinctions facilitated more accurate seg-

mentation (e.g., patients 1 and 3). Despite improvements, challenges persisted, particularly in distinguishing darker areas and incisions. The variation in the results of the 5 models emphasizes the test results' sensitivity to the choice of the test set. Consequently, averaging the test results with k-fold cross validation is a crucial step in order to get a more representative test result for the given dataset.

5 Conclusion

This study showed the effect of pre-training with synthetic data and data augmentation on the liver segmentation in laparoscopic images. The presented results emphasize the need for realistic augmentation strategies. While data augmentation can increase performance, it can also inadvertently decrease it if applied indiscriminately. Additionally, the research showed a clear improvement when applying pre-training on synthetic data in conjunction with fine-tuning on clinical data. Future work involves extending anatomical structure recognition beyond the liver.

Author Statement

This research was funded by Olympus Surgical Technologies Europe. There is no further conflict of interest to declare.

References

- [1] Garry R. Laparoscopic surgery. *Best Practice & Research Clinical Obstetrics & Gynaecology*; 2006;20(1):89-104.
- [2] Hamad GG, Curet M. Minimally invasive surgery. *American Journal of Surgery*; 2010;199(2):263-5.
- [3] Taghanaki SA, et al. Deep Semantic Segmentation of Natural and Medical Images: A Review. *Artificial Intelligence Review*; 2019;54:137-78.
- [4] Bengio Y, Goodfellow I, Courville A. *Deep Learning*. Cambridge, MA, USA: MIT press; 2016.
- [5] Pfeiffer M, et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: *MICCAI*; 2019;22:119-27.
- [6] Ronneberger O, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *MICCAI*; 2015;18:234-41.
- [7] Ma J, et al. Loss odyssey in medical image segmentation. *Medical Image Analysis*; 2021; 71:102035.
- [8] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*. 2014.
- [9] Chorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*; 2019;6(1):1-48.
- [10] Simard PY, et al. Best practices for convolutional neural networks applied to visual document analysis. In: *Icdar*; 2003;3(2003):958-63.