Heiko Walkner*, Lorena Krames, and Werner Nahm

Synthetic Data in Supervised Monocular Depth Estimation of Laparoscopic Liver Images

https://doi.org/10.1515/cdbme-2024-2162

Abstract: Monocular depth estimation is an important topic in minimally invasive surgery, providing valuable information for downstream application, like navigation systems. Deep learning for this task requires high amount of training data for an accurate and robust model. Especially in the medical field acquiring ground truth depth information is rarely possible due to patient security and technical limitations. This problem is being tackled by many approaches including the use of synthetic data. This leads to the question, how well does the synthetic data allow the prediction of depth information on clinical data. To evaluate this, the synthetic data is used to train and optimize a U-Net, including hyperparameter tuning and augmentation. The trained model is then used to predict the depth on clinical image and analyzed in quality, consistency over the same scene, time and color. The results demonstrate that synthetic data sets can be used for training, with an accuracy of over 77% and a RMSE below 10 mm on the synthetic data set, do well on resembling clinical data, but also have limitations due to the complexity of clinical environments. Synthetic data sets are a promising approach allowing monocular depth estimation in fields with otherwise lacking data.

Keywords: Artificial neural networks, Monocular depth estimation, Laparoscopy, Synthetic data

1 Introduction

The transition from open surgery to minimally invasive surgery (MIS) has revolutionized surgical procedures, offering patients numerous benefits including reduced pain, shorter hospital stays, and improved cosmetic outcomes [1]. However, these advancements also bring challenges such as missing tactile feedback, limited field of view, and, in general, more demanding surgeries. To overcome these challenges, programs like navigation assistance systems are being developed [2]. These come with additional demanding tasks, such as 3D to 3D registration, requiring more information about the environment intraoperatively. Since surgeries are performed with monocular or stereo laparoscopes, depth data can be obtained through monocular depth estimation (MDE) for both cases. This study investigates into the potential of MDE through supervised deep learning in complex MIS. However, the current limitation lies in data scarcity due to missing ground truth information needed for a supervised approach. These missing information can not be acquired in a clinical setting due to concerns surrounding patient security and technical limitations. Supervised approaches need annotated data, reducing the viability in fields where the ground truth can not be acquired [3]. In response, synthetic alternatives have emerged, offering a pathway for supervised deep learning. This research aims to explore the benefits and suitability of using synthetic data sets for exclusive training to predict outcomes in clinical settings, using an exemplary data set of photo-realistic laparoscopic images showing the liver region as a proof-of-concept [4, 5]. This analysis aims to understand the role of synthetic data in overcoming data constraints, leading to increased performance and safer information for downstream applications, and contribute insights into the feasibility and utility of synthetic data sets.

2 Methods

The target is to investigate the ability of an artificial neural network (ANN) to generalize on clinical data when only trained on synthetic data. To achieve this the model is trained in a supervised way purely on the synthetic data and tested on both synthetic and clinical data.

2.1 Data Sets

The synthetic data set consists out of 20,000 unique images from 10 different models with corresponding dense depth maps. These images were translated into five different styles, using a generative adversarial network (GAN) on a recreated surgical environment, resulting in 100,000 photo-realistic images with pixel-to-pixel ground truth information allowing for supervised learning [4]. The cholec80 data set, consisting of 80 laparoscopic cholecystectomy surgery videos was used for qualitative evaluation purposes [5]. Images predominantly featuring the liver were manually selected from the data set.

***Corresponding author: Heiko Walkner,** Institute of Biomedical Engineering (IBT), Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, 76131 Karlsruhe, Germany, e-mail: publications@ibt.kit.edu
Lorena Krames, Werner Nahm, IBT, KIT, 76131 Karlsruhe, Germany

2.2 Architecture

In MDE, convolutional neural networks (CNN), recurrent neural networks, and GANs are commonly used architectures. Given the limited data provided by the synthetic data set without time dependency, a U-Net CNN was chosen [6]. U-Nets excel in processing spatial information and are tailored to handle scenarios with limited information, making them well-suited for the given scenario.

2.3 Synthetic Evaluation

The metrics used for quantitative evaluation are the standard for MDE and include the root mean square error (RMSE) and the threshold accuracy metric δ_i, which indicates the percentage of pixels meeting the following condition:

$$\delta_i = \max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < 1.25^i, \tag{1}$$

where d represents each pixel's estimated depth, d^* denotes the corresponding ground truth depth and $i \in \{1, 2, 3\}$ for 3 thresholds, as suggested by previous works [7]. The first part of the study focuses on optimizing the network to allow for an accurate and robust performance on the synthetic data set and generalization for the clinical data set, using 8 models of the data set for training and 1 for validation and test each. The optimization focuses on hyperparameter tuning in form of comparing loss functions and learning rate schedulers. The compared loss functions are L1, L2 and BerHu loss, which other research has shown to be the most suitable for MDE [8]. Three different approaches for learning rate scheduler were evaluated: static, dynamic, and cyclic learning rates. To improve the generalization the input images were augmented by changing the image colors and cropping the images. The color augmentation was tested in three different intensities: none, middle and strong augmentation. The brightness, contrast, and saturation were adjusted by 0, 0.5 and 0.5 from 1 and the hue by 0, 0.03, 0.1 from 0.5. The results are compared by utilizing the five styles provided by the data set. Since these styles represent a form of unknown augmentation, four of them were employed for both training and testing to assess overall accuracy. The remaining style was exclusively reserved for testing to evaluate performance on unseen augmentation.

2.4 Clinical Evaluation

To evaluate predictions on the clinical data set, the best performing model from Sec. 2.3 was employed. The missing ground truth for the clinical data set only allows for a qualitative evaluation, which is divided into three main steps. The first step involved translating frames from the clinical data set into depth maps and analyzing them for inconsistencies to provide a general qualitative evaluation. Do the proportions and trends in the predictions match the expected results? In the second step, the impact of color differences between the clinical and synthetic data sets on the results was examined, investigating if color augmentation, due to differences in optics, alter the predictions. This was achieved by adjusting the RGB-values of the clinical data to match the mean and standard deviation of the synthetic data set, and then comparing the results with the original ones. The third step evaluated both spatial and temporal consistency, evaluating if the network is consistent and where differences occur. Spatial consistency was assessed by dividing images into parts for separate prediction, while temporal consistency was examined by comparing predictions across successive frames.

3 Results

3.1 Synthetic Evaluation

The analysis of hyperparameter tuning examined the impact of different loss functions and learning rate schedulers. Results revealed minimal divergence in accuracy metrics across varied loss functions as shown in Table 1. For the learning rate schedulers, the dynamic one resulted in the best average accuracy and RMSE. Investigating the influence of augmentation levels on model efficacy showcased that while the most robust augmentation led to the highest RMSE, the absence of augmentation yielded the best RMSE score. However, middle-level augmentation demonstrated superior accuracy metrics. Furthermore, when compared on an unseen style, middle-level augmentation surpassed no augmentation across all metrics as seen on Table 2. These experiments resulted in the best achieved model in Table 3 showing results like Figure 1.

Tab. 1: Results of the model being trained on synthetic data, comparing different loss functions and learning rate schedulers.

Loss Function	$\delta_3 \uparrow$	$\delta_2 \uparrow$	$\delta_1 \uparrow$	**RMSE**$\downarrow$
L1	98.03%	93.32%	68.94%	8.28 mm
L2	98.04%	92.85%	67.56%	8.21 mm
BerHu	98.12%	93.08%	67.90%	8.38 mm
Learning Rate	$\delta_3 \uparrow$	$\delta_2 \uparrow$	$\delta_1 \uparrow$	**RMSE**$\downarrow$
Static	98.03%	93.32%	68.94%	8.28 mm
Dynamic	98.27%	93.37%	70.19%	8.02 mm
Cycle	98.10%	92.68%	69.13%	8.40 mm

3.2 Clinical Experiment

The translation of clinical images is shown Figure 2. In this image, the anticipated lowest depth was observed at the bottom, corresponding to the location of the liver, with depth increasing towards the abdominal wall in the background. Moreover, the ligament on the liver is depicted within the anticipated

Tab. 2: Results of the model being trained on four styles the synthetic data with different augmentations. First tested on the four styles to compare the different augmentations. Then tested on the unseen style five to estimate the generalization of the network.

Augmentation	$\delta_3\uparrow$	$\delta_2\uparrow$	$\delta_1\uparrow$	**RMSE**$\downarrow$
No	97.16%	91.25%	67.14%	10.86 mm
Middle	97.44%	91.37%	67.80%	11.31 mm
Strong	97.56%	90.80%	64.53%	12.05 mm
Unseen Style	$\delta_3\uparrow$	$\delta_2\uparrow$	$\delta_1\uparrow$	**RMSE**$\downarrow$
No	96.97%	89.29%	62.43%	19.82 mm
Middle	98.30%	91.86%	68.01%	10.89 mm

Tab. 3: Results of the model being trained on synthetic data with augmentation.

Metric	$\delta_3\uparrow$	$\delta_2\uparrow$	$\delta_1\uparrow$	**RMSE**$\downarrow$
Result	99.19%	96.22%	77.15%	9.98 mm

depth range and is distinctly separated from the abdominal wall. In images portraying scenes not present in the synthetic data set, errors may be observed. This is shown in Figure 3, where the puncture of the abdominal wall, expected to be a low depth, is not accurately depicted in the prediction. In the second step, adjusting the color to match the mean and standard deviation of the RGB-values, demonstrated minimal differences in predictions. Figure 4 displays the resulting predictions and their variances, with only portions of the background showing slightly greater disparities between the images. Spatial consistency assessments, achieved through image segmentation for separate prediction comparisons, showcased precise predictions across various image regions. Only in images with little information or light, as seen in the top left corner of Figure 5, the predictions differ over multiple images. Furthermore, temporal consistency examinations, involving the comparison of predictions across successive frames, underscored the model's stability over time. These findings jointly demonstrated the robustness and accuracy of the model's predictions, underscoring their potential for clinical applications.

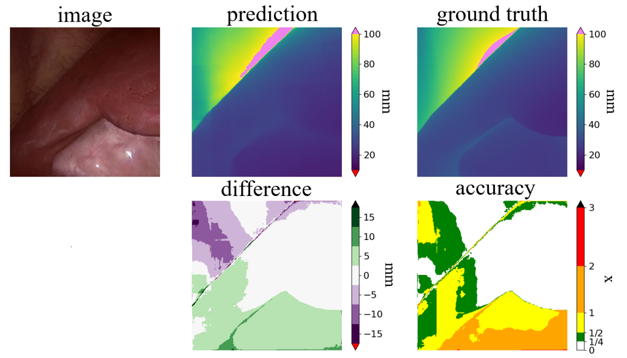

Fig. 1: Synthetic data: Image, prediction, ground truth information as well as the metrics in form of the difference and accuracy.

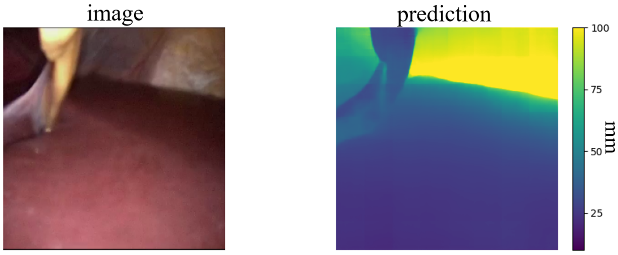

Fig. 2: Clinical image on the left and the depth prediction on the right. The image depicts the liver with the ligament in the front and the abdominal wall in the background.

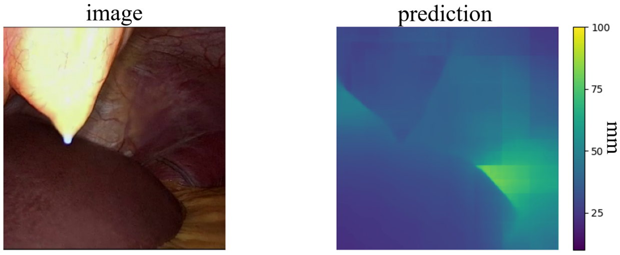

Fig. 3: Clinical image on the left side and the depth prediction on the right side. The image depicts a puncture of the abdominal wall in the front, with the liver and the abdominal wall in the back.

4 Discussion

4.1 Synthetic Evaluation

The network seems to be capable of estimating depth within the synthetic data mimicking photo-realistic laparoscopic scenes. The experiments indicate that all loss functions yield similarly favorable outcomes, with each excelling in specific metrics. L1 loss was chosen due to δ_1, showing the most

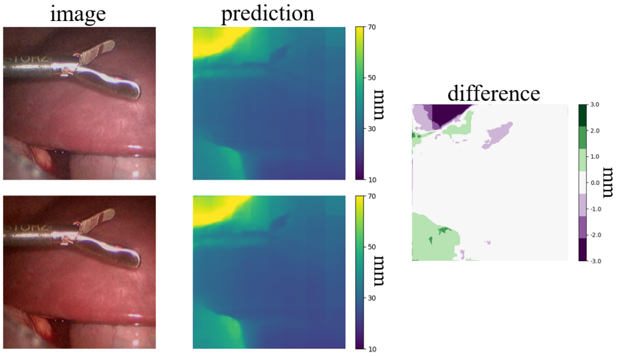

Fig. 4: Original image with the depth prediction in the first row, and the color adjusted image and the corresponding translation in the second row. On the right, the difference calculated by subtracting the original prediction from the color adjusted prediction.

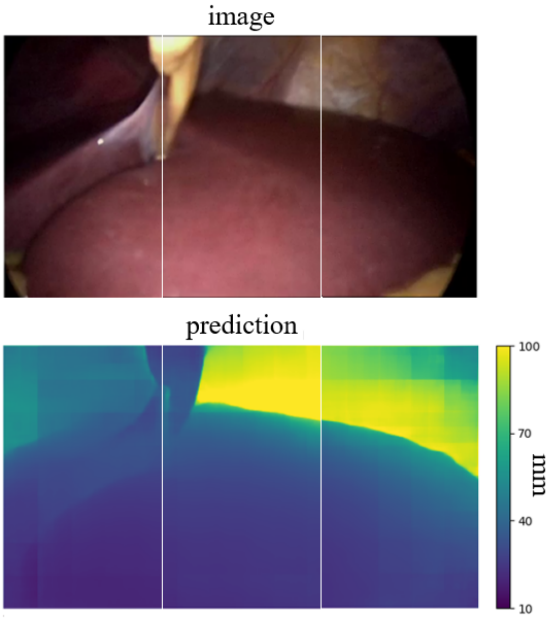

Fig. 5: Image split into three parts with corresponding depth predictions.

promising results with focus on the liver. RMSE and δ_3 are more affected by the background. Regarding learning rate schedulers, the dynamic approach demonstrates superior performance. The method of enhancing hyperparameters may result in suboptimal outcomes due to parameter interdependence. A method like grid search could lead to potential improvements. The augmentation results indicate that the middle configuration preserves metrics and enables generalization across different optics or styles. Specifically, the first part demonstrates that the middle augmentation maintains accuracy on known styles, while the second part reveals a significant drop in accuracy over unknown styles if no augmentation is utilized.

4.2 Clinical Evaluation

The translations of clinical data indicated that scenes similar to the synthetic data set produced expected results, while unfamiliar scenes resulted in errors. The post-training color augmentation of clinical images reveals that color adjustment does not significantly alter predictions in the areas of interests. These findings suggest that color adjustment can be compensated by simple augmentation in the training. The spatial consistency showed only minor inconsistencies in depth predictions across overlapping regions were observed, maintaining a consistent form. In areas with little information, through cut of of scenes, capturing more images from diverse viewpoints and gathering additional scene information could potentially mitigate such inconsistencies. The temporal consistency of depth predictions across sequential frames indicates consistent depth and form across frames, although factors like reflections can introduce deviations, if the reflections cover information like boarders between organs. Future solutions may involve enabling the ANN to gather more information, such as employing time-sensitive inputs to incorporate data from previous frames for more accurate predictions.

5 Conclusion

This study demonstrates the potential of synthetic data sets in addressing data scarcity challenges in MDE for MIS. By leveraging supervised learning with U-Net architecture and conducting optimization and generalization experiments, we have demonstrated the feasibility of training deep learning models exclusively on synthetic data for clinical outcome predictions. Our findings indicate the effectiveness of synthetic data sets in yielding promising results with high accuracy metrics and minimal RMSE on synthetic data. However, transferring this performance to clinical data remains a complex task, requiring careful consideration of various factors such as data set complexity, necessitating consideration of factors like data set complexity. Nevertheless, synthetic data sets offer a controlled and privacy-compliant approach to data generation, facilitating safer and more efficient training of deep learning models in the medical field. Further research is needed to address the remaining challenges and enhance the generalization capabilities of models trained on synthetic data for real-world clinical applications. Especially, a quantitative evaluation of the clinical data is necessary.

Author Statement

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

References

[1] Morise Z., Current status of minimally invasive liver surgery for cancers. World J Gastroenterol 2022; 28(43): 6090–6098.
[2] Schneider C., et al., Performance of image guided navigation in laparoscopic liver surgery – A systematic review. Surg. Oncol., vol. 38, 101637, 2021.
[3] Ming Y., et al., Deep learning for monocular depth estimation: A review. Neurocomputing, vol. 438, pp. 14–33, 2021.
[4] Pfeiffer M., et al., Generating Large Labeled Data Sets for Laparoscopic Image Processing Tasks Using Unpaired Image-to-Image Translation. MICCAI 2019.
[5] Twinanda A.P., et al., EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos, IEEE TMI, 2017
[6] Schilling T., et al., Engineering of beiodegradable magnesium alloy scaffolds to stabilize biological myocardial grafts. Biomed Eng-Biomed Tech 2017;62:493–504.
[7] Khan F., et al., Deep Learning-Based Monocular Depth Estimation Methods—A State-of-the-Art Review. 2020 Sensors (Basel), 2020, 20(8), 2272
[8] Carvalho M., et al., On Regression Losses for Deep Depth Estimation. 2018 25th IEEE ICIP, 2018, pp. 2915-2919