Medical Image Analysis (2024)

Contents lists available at ScienceDirect

Medical Image Analysis

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

An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion from the MICCAI2022 challenge

Sharib Ali[a,n,*], Yamid Espinel[b,n], Yueming Jin[e], Peng Liu[g], Bianca Güttner[g], Xukun Zhang[h], Lihua Zhang[h], Tom Dowrick[i], Matthew J. Clarkson[i], Shiting Xiao[j], Yifan Wu[k], Yijun Yang[l], Lei Zhu[l], Dai Sun[m], Lan Li[m], Micha Pfeiffer[g], Shahid Farid[d], Lena Maier-Hein[f], Emmanuel Buc[b,c], Adrien Bartoli[b,c]

[a] *School of Computing, Faculty of Engineering and Physical Sciences, University of Leeds, Leeds, LS2 9JT, United Kingdom*
[b] *Centre Hospitalier Universitaire de Clermont-Ferrand, Clermont-Ferrand, France*
[c] *Endoscopy and Computer Vision Group, Université Clermont Auvergne, Clermont-Ferrand, France*
[d] *Department of HPB and Transplant Surgery, St. James's University Hospital, Leeds, United Kingdom*
[e] *Department of Electrical and Computer Engineering National University of Singapore (NUS), 119276, Singapore*
[f] *German Cancer Research Center (DKFZ), Heidelberg, Germany*
[g] *Department of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC Dresden), Fetscherstraße 74, 01307 Dresden*
[h] *Academy for Engineering and Technology, Fudan University, Shanghai, China*
[i] *Wellcome EPSRC Centre for Interventional and Surgical Sciences, UCL, London, UK*
[j] *Department of Computer and Information Science, University of Pennsylvania, PA 19104, Philadelphia, US*
[k] *Department of Bioengineering, University of Pennsylvania, PA 19104, Philadelphia, US*
[l] *Hong Kong University of Science and Technology (Guangzhou), 511455, Guangzhou, China*
[m] *Suzhou Institute for Advanced Research, Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE), University of Science and Technology of China, Suzhou, China*
[n] *these authors contributed equally to this work*

ARTICLE INFO

Article history:

ABSTRACT

Augmented reality for laparoscopic liver resection is a visualisation mode that allows a surgeon to localise tumours and vessels embedded within the liver by projecting them on top of a laparoscopic image. Preoperative 3D models extracted from Computed Tomography (CT) or Magnetic Resonance (MR) imaging data are registered to the intraoperative laparoscopic images during this process. Regarding 3D-2D fusion, most algorithms use anatomical landmarks to guide registration, such as the liver's inferior ridge, the falciform ligament, and the occluding contours. These are usually marked by hand in both the laparoscopic image and the 3D model, which is time-consuming and prone to error. Therefore, there is a need to automate this process so that augmented reality can be used effectively in the operating room. We present the Preoperative-to-Intraoperative Laparoscopic Fusion challenge (P2ILF), held during the Medical Image Computing and Computer Assisted Intervention (MICCAI 2022) conference, which investigates the possibilities of detecting these landmarks automatically and using them in registration. The challenge was divided into two tasks: 1) A 2D and 3D landmark segmentation task and 2) a 3D-2D registration task. The teams were provided with training data consisting of 167 laparoscopic images and 9 preoperative 3D models from 9 patients, with the corresponding 2D and 3D landmark annotations. A total of 6 teams from 4 countries participated in the challenge, whose results were assessed for each task independently. All the teams proposed deep learning-based methods for the 2D and 3D landmark segmentation tasks and differentiable rendering-based methods for the registration task. The proposed methods were evaluated on 16 test images and 2 preoperative 3D models from 2 patients. In Task 1, the teams were able to segment most of the 2D landmarks, while the 3D landmarks showed to be more challenging to segment. In Task 2, only one team obtained acceptable qualitative and quantitative registration results. Based on the experimental outcomes, we propose three key hypotheses that determine current limitations and future directions for research in this domain.

Introduction

Laparoscopic liver resection (LLR) is a minimally invasive procedure used in the removal of benign or malignant tumours. It has become increasingly popular in the last two decades owing to the reduced trauma to the patient and the shorter hospitalisation times. However, it remains a challenging technique due to the reduced intra-abdominal space and the lack of tactile feedback. This makes it difficult to find intraparenchymal structures like tumours and vessels, which increases the risk of wrong resections. Augmented Reality (AR) could mitigate this issue by overlaying a 3D model reconstructed from Computed Tomography (CT) or Magnetic Resonance (MR) imaging onto the laparoscopic views, as shown in Figure 1. Only one of both modalities is required to reconstruct the 3D models, provided the desired structures are clearly visible. As depicted, the surgeons can then see the inner structures, and perform tumour resection accordingly. Owing to the liver's substantial flexibility, a deformable registration should be done to fit the preoperative 3D model with the intraoperative data effectively. Once the registration is computed, the fusion can be realised. In terms of registration accuracy and according to Zhong et al. (2017), a margin of healthy tissue around the tumour of at least 1 cm should be kept for Hepatocellular Carcinoma (HCC) resections. Therefore, an AR system can be considered to be effective if its target registration error (TRE) is lower than 1 cm.

Registration for Augmented Reality

Existing methods register the 3D preoperative data into 3D or 2D intraoperative data. Most of these methods use liver anatomical landmarks to constrain registration and help the preoperative model to fit in the intraoperative data. For the 3D-3D registration case, some examples are found in (Robu et al., 2018; Modrzejewski et al., 2019), where the landmarks are marked manually on both the preoperative and intraoperative 3D shapes. The main problem of the 3D-3D registration methods is that they reconstruct the intraoperative data from stereoscopic cameras, which are not always available in surgery rooms. They may also use 3D reconstruction algorithms like Structure-from-Motion (SfM) or Simultaneous Localisation and Mapping (SLAM), which only work in rigid scenes and generally fail for the non-rigid liver. For the 3D-2D registration case, some examples are found in (Adagolodjo et al., 2017; Koo et al., 2017; Espinel et al., 2022; Koo et al., 2022; Labrunie et al., 2022), where the landmarks are marked in the intraoperative images either manually or automatically, but always marked manually on the preoperative 3D models. In this work, we focus on the 3D preoperative to 2D laparoscopic image registration problem, which aligns the preoperative 3D models to one or several intraoperative 2D images.

According to (Koo et al., 2017; Espinel et al., 2022), some of the landmarks that can be used in 3D preoperative to 2D laparoscopic image registration are the liver's lifted *ridge*, the *falciform ligament*, and the *silhouette*, as shown in Figure 2. The *ridge* landmark corresponds to the pronounced curve located at the bottom-anterior part of the liver. It is the most distinguishable landmark among the three and covers both the left and right lobes of the liver. The *falciform ligament* is the thin, fibrous tissue that connects the anterior part of the liver to the abdominal wall. It is usually cut during a laparoscopic procedure to facilitate the manipulation of the liver. The remnant of this tissue on the liver's surface is what we use as a landmark. The *silhouette* landmark corresponds to the boundary of the liver at a given image and, thus, does not have a direct correspondence in the 3D model. The 3D correspondences are usually found during registration using algorithms like the Iterative Closest Point (ICP). In order to accurately fit the 3D model to the images, a good correspondence between the landmarks in the laparoscopic image and the preoperative 3D model should be found. However, as the marking is usually done by hand, it will greatly depend on the user's understanding of the scene, which can be a source of inaccuracies. Moreover, the time required to manually mark these landmarks, usually several minutes, makes it difficult to integrate AR within the surgical workflow. Some of the existing methods segment the landmarks on the images automatically like the works in (Labrunie et al., 2022; Koo et al., 2022), where the 2D liver landmarks are segmented using deep learning, but the 3D landmarks are still marked manually. Due to the limitations above of manual marking in terms of accuracy and time, there is a need for automating the segmentation of these landmarks in the image and the preoperative 3D models, as well as accurately finding the correspondences between them for the 3D preoperative to 2D laparoscopic image registration.

Presentation of the challenge

The Preoperative-to-Intraoperative Laparoscopic Fusion (P2ILF) challenge addresses the problem of finding the liver's anatomical landmarks in both the laparoscopic images and the preoperative 3D model, and of using them for 3D preoperative to 2D laparoscopic image registration. This challenge was deployed on the Grand Challenge platform (Ali et al., 2022), where the teams could register, download the training data, upload their algorithms, and run them on the test data. The challenge was divided into two phases. In phase I of the challenge, the participants had to segment the visible 2D landmarks in the laparoscopic images, and then segment the corresponding 3D landmarks in the preoperative 3D model. In phase II of the challenge, the participants had to perform 3D preoperative to 2D laparoscopic image registration. For phase II, the participants were suggested to use the 3D and 2D landmarks segmented in phase I. However, this was not mandatory, and they could perform either a rigid or a deformable registration. For this challenge, surgical data was collected and annotated for 11 patients, including their corresponding preoperative 3D models, the intraoperative laparoscopic images, and the intrinsic camera parameters. The provided data presents two main challenges: the drastic change in shape and appearance of the liver between patients and the limited amount of data. A total of six teams from four countries participated in the challenge. We describe the algorithm developed by each team and the results obtained on the test set.

*Corresponding author
e-mail: s.s.ali@leeds.ac.uk (Sharib Ali)

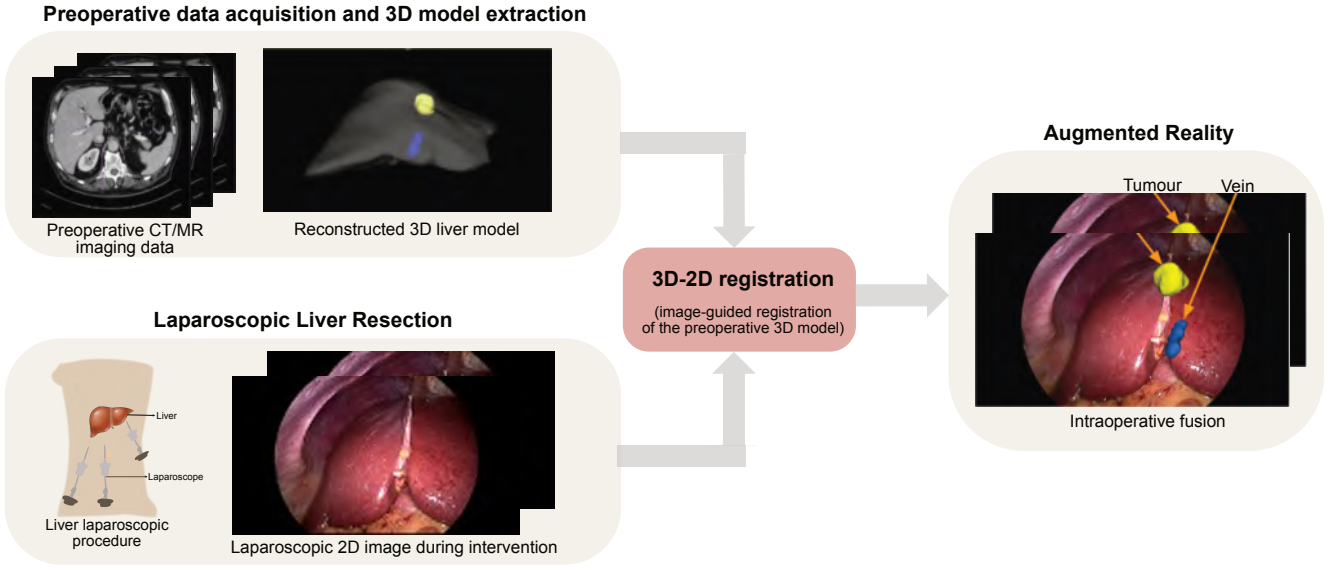

Fig. 1: **Laparoscopic image fusion with preoperative 3D CT or MR scans.** A preoperative 3D scan is first used to reconstruct the liver boundaries, tumours and major vessels critical for a safe surgery. During the laparoscopic procedure we overlay the reconstructed model using image registration, in this case 3D meshes, to the 2D liver view. The idea is to project 3D mesh points onto the liver boundaries that can enable understanding of the spatial location of the tumours and vessels along with the matched liver boundaries in the acquired 3D model. Such an augmented reality technique helps surgeons to locate the tumour and important landmarks during surgery. The above results were obtained with the semi-automatic method from Koo et al. (2017).

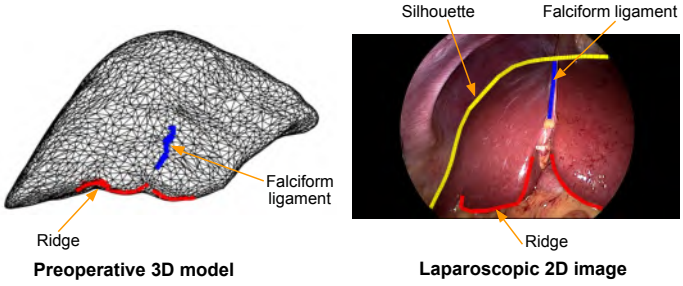

Fig. 2: **Depiction of the 2D and 3D anatomical landmarks.** Anatomical liver landmark ground-truth annotations in the preoperative 3D model (left), and in the laparoscopic 2D image (right).

In this paper, we first present the related work, the details on the newly curated dataset for AR in LLR, the design and setup of the challenge in more detail, the methods proposed by the participating teams, results and insights regarding the limitations of each approach, and finally we conclude with the discussions presenting empirical and experimental hypotheses and future work.

Related work

Existing datasets for AR

Currently, there is a lack of publicly available datasets for AR, and most of the existing AR methods use non-publicly available data, dealing only with the 2D landmark automatic segmentation problem. For example, Koo et al. (2022) used a private dataset of 133 images coming from two patients to segment the anatomical landmarks of the liver (including the *ridge* and the *silhouette*). The method heavily relied on synthetic data generation. Labrunie et al. (2022) used a dataset of 1415 laparoscopic images coming from 68 patients to segment the liver landmarks, but their dataset is not publicly available either. There are some available datasets containing endoscopic liver videos like the Cholec80 dataset (Twinanda et al., 2017), HeiChole dataset (Wagner et al., 2023), and the Dresden Surgical Anatomy Dataset (Carstens et al., 2023), but they do not contain the preoperative data and the intrinsic camera parameters required to work with AR.

Registration for AR in LLR has been an active field of research over the last decade, with the existing methods using either monocular endoscopes, stereo endoscopes, and external devices like optical trackers and intraoperative CT scanners. These methods can be globally classified into 3D-2D and 3D-3D registration methods, if the preoperative 3D model is registered to an intraoperative 2D image or an intraoperative 3D model.

3D-2D registration methods

A single-view monocular method is presented by Koo et al. (2017), which we use as basis and motivation of our work. In this work, the authors combined the *ridge*, *falciform ligament*, and *silhouette* landmarks with a biomechanical model to perform registration. Prior to registration, the landmarks are manually marked in both the 2D image and the preoperative 3D model. It uses a Gauss-Seidel iterative algorithm to solve the landmark and biomechanical constraints. Other monocular 3D-2D registration methods use one or multiple images simultaneously. For example, a set of *silhouette* landmarks were manually marked in the image and combined with a biomechanical model to drive registration by Adagolodjo et al. (2017). These constraints were solved using a Gauss-Seidel iterative optimisation approach. A set of methods that perform 3D preoperative to 2D laparoscopic image registration on multiple laparoscopic images is presented by Espinel et al. (2022), where the

anatomical landmarks from all the images are combined to deal with the partial visibility problem and improve registration accuracy. In this case, the landmarks should be manually marked on each image separately, which increases the total registration time. In an attempt to reduce the risk of wrong annotations and the registration time, the method by Koo et al. (2022) segments the anatomical landmarks in the images automatically. To achieve this, a CASENet CNN was trained with a small dataset of 133 patient images, along with a synthetic dataset consisting of 100,000 images. However, the 3D landmarks were still marked manually in the 3D model. The proposed rigid registration starts by computing a canonical liver pose, assuming that the camera is inserted close to the belly button. Then, a set of transformations was generated by randomly rotating the model about the three axes. For each of the transformations, the closest points between the 3D and 2D landmarks are found, and an optimal transformation is estimated using Perspective-*n*-Point (P*n*P) with RANSAC. In the end, the best transformation is chosen based on the minimum Hausdorff distance between the 3D and 2D landmarks. Similarly, another approach that segments the landmarks automatically was proposed by Labrunie et al. (2022), where an off-the-shelf UNet was trained with 1415 laparoscopic images from 68 patients. Again, the 3D model landmarks were still annotated manually before surgery. The main goal of this work was to perform an initial rigid registration to serve as the basis for subsequent deformation stages. The registration approach used a RANSAC-based P*n*P strategy that iteratively recomputed the correspondences between the 2D and 3D landmarks.

3D-3D registration methods

Some monocular methods may perform 3D-3D registration like the one presented by Modrzejewski et al. (2019), where the shape of the liver was reconstructed using SfM during the intraoperative procedure. This shape was then combined with a set of landmarks and a biomechanical model to perform deformable registration. The registration process follows a rigid-to-deformable energy minimisation strategy, which runs until the convergence threshold is reached. Another method that uses SfM is presented by Cheema et al. (2019), where an intraoperative shape also serves as a target for registration. In this case, correspondences between the preoperative and the intraoperative shapes were combined with shading cues to align and deform the intraoperative shape. Similarly, Espinel et al. (2022) combined the reconstructed camera poses with the anatomical landmarks and the biomechanical parameters for registration. However, Espinel et al. (2022) suggested that applying SfM in liver scenes is difficult due to the constant deformations and the limited range of camera movements. Methods that use stereoscopic cameras or other external devices usually perform 3D-3D registration. In particular, the methods by Haouchine et al. (2013); Soler et al. (2014); Thompson et al. (2015); Bernhardt et al. (2016); Robu et al. (2018); Luo et al. (2020) reconstruct an intraoperative 3D model of the visible liver using stereoscopic techniques. In these cases, the intraoperative 3D model is used as a target to register the preoperative 3D models. Some of the methods perform rigid registration (Soler et al., 2014; Thompson et al., 2015; Bernhardt et al., 2016; Robu et al., 2018; Luo et al., 2020), while the method by Haouchine et al. (2013) is the only work that performs deformable registration. In addition to a stereo endoscope, the method from Thompson et al. (2015) also uses an optical tracker to locate and merge multiple stereoscopically reconstructed patches of the intraoperative liver. A major limitation of these methods is the requirement of stereo endoscopes and external tracking devices that are not commonly available in surgery rooms.

In this work, we aim to find registration methods that only use the available preoperative models and a monocular endoscopic setting in the surgery room. Such methods should automatically find the liver anatomical landmarks that can then lead to computing 3D preoperative to 2D laparoscopic image registration automatically. We attempt to motivate the usage of data-driven approaches, which is still uncommon in this problem, as well as to reduce both the user interactions and the registration times. Given the high number of existing methods and a lack of unified comparison, this challenge is the first one to provide an objective comparison between registration methods for AR in LLR, which is a requirement to continue advancing in the field.

The P2ILF challenge

General aspects of the dataset

The training dataset is composed of 9 patients with 167 laparoscopic images, their corresponding 2D and 3D anatomical landmarks, their respective preoperative 3D models, and the intrinsic camera parameters. The test dataset is composed of 2 patients and includes 16 selected images (8 images per patient) with their corresponding preoperative 3D models and the intrinsic camera parameters. A quantitative description of the whole dataset is given in Table 1. It includes the number of images per patient, the type of preoperative images (CT/MR) used to reconstruct the preoperative 3D models, and the liver condition (cirrhotic/non-cirrhotic). The training dataset was provided to the participants, who were allowed to freely split the data for training and validation. However, for the test phase, an online platform was used in a way that prohibited the teams from accessing the test samples directly. For the algorithmic testing reported in this paper, each algorithm was evaluated through the deployment of Docker containers.

Figure 3 illustrates the training and test samples for the 11 patients (with one sample per patient), including the original laparoscopic images at the top, the ground-truth anatomical landmarks in the middle (with the *silhouette*, *ridge* and *falciform ligament* in yellow, red, and blue, respectively), and the preoperative 3D models with their corresponding 3D landmarks for the *ridge* (in red) and the *falciform ligament* (in blue). It can be observed that the appearance of the liver varies greatly across patients. Moreover, some of the patients have a visible ultrasound probe, which is common in laparoscopy as it may be used to identify key vessels and tumour locations during surgery. To better evaluate the generalisation of the proposed methods, we used one patient with cirrhotic liver and one patient with non-cirrhotic liver in the test dataset.

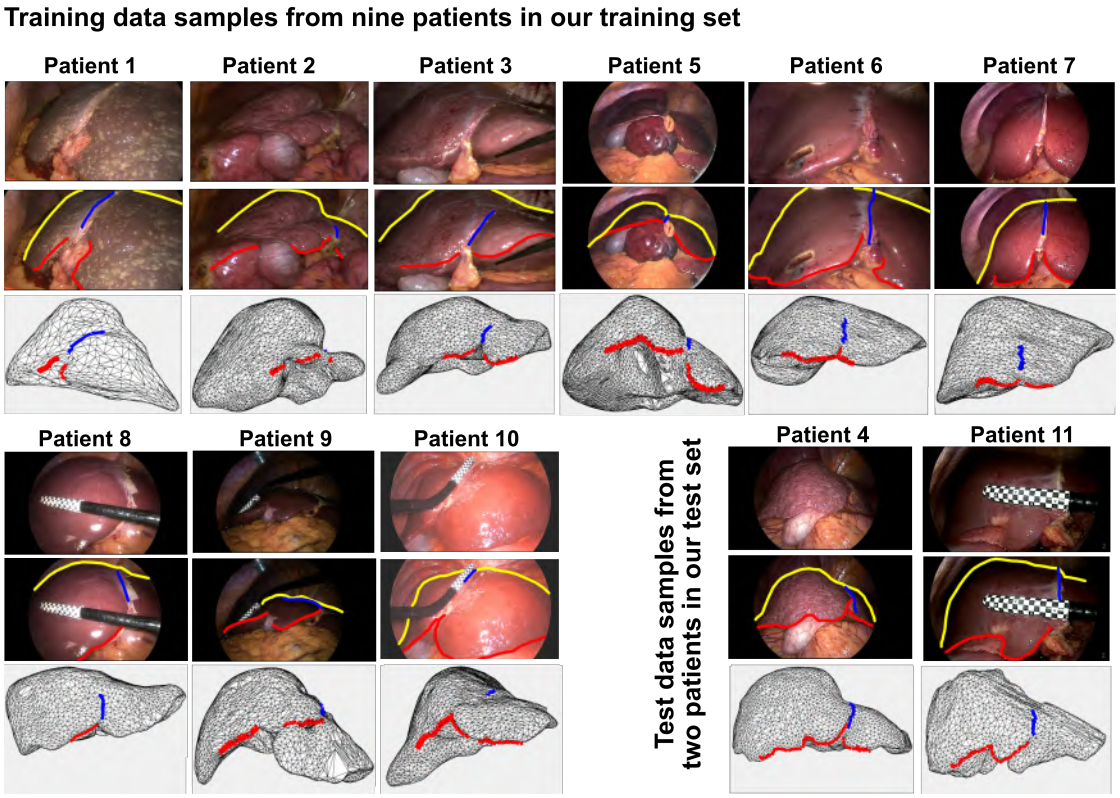

Fig. 3: **P2ILF dataset:** Training and test data samples with original laparoscopic images, annotated anatomical landmarks (*silhouette* in yellow, *ridge* in red and *falciform ligament* in blue), and the corresponding 3D anatomical annotations (*rigde* in red and *falciform ligament* in blue) in manually aligned 3D liver models are provided. The dataset contains a total of 11 patients, divided in 9 patients for training and 2 patients for testing.

Table 1: Quantitative description of the generated patient dataset for the P2ILF Challenge

Patient #	Preoperative imaging type	# Intraop. images	Liver condition	Data type
1	CT	22	C	TR
2	CT	25	C	TR
3	CT	20	NC	TR
4	MR	8	C	TS
5	CT	15	NC	TR
6	CT	22	NC	TR
7	MR	25	NC	TR
8	CT	8	NC	TR
9	MR	21	NC	TR
10	CT	9	NC	TR
11	CT	8	NC	TS
Total:	**8 CT, 3 MR**	**183**	**3 C, 8 NC**	**9 TR, 2 TS**

Intraop.: Intraoperative; CT: Computed Tomography; MR: Magnetic Resonance; C: Cirrhotic; NC: Non-cirrhotic; TR: Training; TS: Testing

Ethical and privacy aspects of the data

The preoperative and intraoperative data of this dataset were collected from the University Hospital of Clermont-Ferrand, France. The data collection was supported by an ethical approval with ID IRB00008526-2019-CE58 issued by CPP Sud-Est VI in Clermont-Ferrand, France. Patient consent to record data was obtained before each intervention. The intraoperative video streams were captured using laparoscopic cameras. All the collected data were fully anonymised before publication. In other words, no meta information (name, birth date, gender, etc.) was passed to the participating teams. During the challenge, all participants were required to sign a data privacy statement. Redistribution or transfer of the data was strictly prohibited. The data upon public release will be free to use (under licence CC-by-NC-SA 4.0) after the publication.

Video collection and dataset construction

The dataset for the P2ILF challenge consists of two types of data: preoperative 3D liver models and intraoperative 2D laparoscopic images. The data were collected by the following procedure:

- Several days before the liver surgery, 3D CT/MR images of the patients were obtained. The liver, the tumours and the vena cava were manually segmented in the CT/MR images by an experienced hepatobiliary surgeon using MITK (German Cancer Research Center (DKFZ), 2008). The surgeon first segmented the liver in every slice using a combination of a region-growing tool with a manual selection tool. Then, they segmented the tumours and the vena cava using the manual selection tool. The brightness and contrast of the images were varied in some cases to improve the visibility of the structures. After the structures were segmented, a 3D interpolation was made between the 2D masks of each structure to generate the 3D models.

- During each surgery, an exploration of the intra-abdominal

scene was done in such a way that the liver was visible to the camera. A video was captured during the exploration. To estimate the intrinsic camera parameters, a video of a moving checkerboard pattern was also captured with the laparoscope.

- To estimate the camera intrinsic parameters, images were first extracted from the checkerboard video at a rate of 5 frames per second, ensuring a sufficient movement of the checkerboard between images. From this frame set, 30 to 40 images where the checkerboard is sharp enough were selected, meaning the corners and edges were distinguishable. Finally, the images were imported into the Metashape software (Agisoft LLC, 2023) and the intrinsic camera parameters were estimated. These parameters included the camera's focal length, the principal point, and the lens distortion.

From the raw laparoscopic videos, the laparoscopic images were selected based on two criteria:

- Noticeable viewpoint change between the images: To ensure a sufficient camera displacement, images were extracted from the laparoscopic videos at a rate of 5 frames per second.

- Clear sharpness in terms of focusing and blur: To ensure a good image quality, images with a clear separation between the liver and the surrounding structures were selected from the previously extracted set. This was done visually by the challenge organisers, with the assistance of an expert surgeon.

Annotation strategies and quality assurance

Due to a lack of available LLR datasets with annotated 2D/3D anatomical landmarks, there was a need to annotate the landmarks in multiple images and preoperative 3D models. To achieve this, three of the challenge organizers, guided by the indications given by two hepatobiliary surgeons with over 10 years practicing experience and one computer scientist with over 5 years experience in working in AR for LLR, proceeded to annotate the 2D landmarks in the 183 laparoscopic images and the corresponding 3D landmarks in the 11 preoperative 3D models. All the annotators have worked extensively in artificial intelligence for surgical image analysis for over 5 years. Each annotator labelled a specific set of images, with every image being annotated only once. The annotations were first reviewed by the scientists and then reviewed together with the surgeons. The labels were corrected where necessary, according to the feedback from the surgeons. It is to be noted that preoperative 3D model annotations were done by the scientist together with the surgeons, due to the complexity in identifying the landmarks.

The annotators were required to annotate the *ridge*, the *silhouette*, the *falciform ligament*, and the liver surface in every image. The tolerance error of annotation was 5 pixels. If the distance between the annotation and the actual landmark exceeded this tolerance range, it was rejected in a later review. All the annotations were done via LabelBox, an open-source collaborative web-based tool. For the annotations to be as precise as possible, the annotators were advised to use annotation tablets to perform their tasks. Some important protocols that were agreed upon and communicated for the annotation process were:

- The *falciform ligament* is on the liver surface and should divide the right and left lobes
- The *ridge* is the curvy area located at the bottom of the liver's posterior part
- The *silhouette* is the occluding boundary of the liver, usually located at the upper part of the liver
- The *silhouette* should not go inward the *falciform ligament* margin, but rather go over it
- The landmarks occluded by blood, neighbouring organs, or surgical tools should not be considered

Challenge tasks and setup

We evaluated the teams on the following two tasks - a) **Task 1:** We requested the teams to perform 2D landmark segmentation on the laparoscopic images and 3D landmark segmentation on the preoperative 3D models as two sub-tasks. Landmark segmentation in the 2D laparoscopic images and the 3D preoperative models are tackled under the same task since the 2D and 3D landmarks should eventually provide correspondences for later registration. Hence, it is important to allow both modalities to be segmented jointly in order for the segmentation to embody the notion of corresponding landmarks. For the 2D case, the teams were asked to segment the *ridge*, the *silhouette* and the *falciform ligament* landmarks. For the 3D case, they were asked to segment the *ridge* and the *falciform ligament* landmarks, according to the previously segmented 2D landmarks. We provided the teams with the 167 laparoscopic images, the camera calibration parameters for each patient, the 9 preoperative 3D models, and the corresponding 3D-2D landmark annotations from the 9 training patients. We kept the 2 test patients undisclosed and used their 3D-2D landmark annotations as ground truth to assess the predictions done by the teams. However, as the proposed methods by the teams segment the 2D and 3D landmarks independently so, for clarity within task 1, we have referred them as two sub-tasks.

b) **Task 2:** We requested the teams to register the preoperative 3D models into the intraoperative laparoscopic images, preferably by using the previously predicted 2D and 3D landmarks. This 3D preoperative to 2D laparoscopic image registration could be either rigid or deformable. We used the 2D *ridge*, the *falciform ligament*, and the *silhouette* landmarks from the 2 test patients as groundtruth to assess the registrations done by the teams.

The input and output data to be used in each of the tasks are shown in Figure 4. The teams were required to run their methods in a Docker-based deployment framework, hosted in the Grand Challenge platform (Radboud University Medical Center, 2023). The test data was not accessible by the teams. Another Docker-based container was developed to assess the predictions and generate the evaluation metrics automatically.

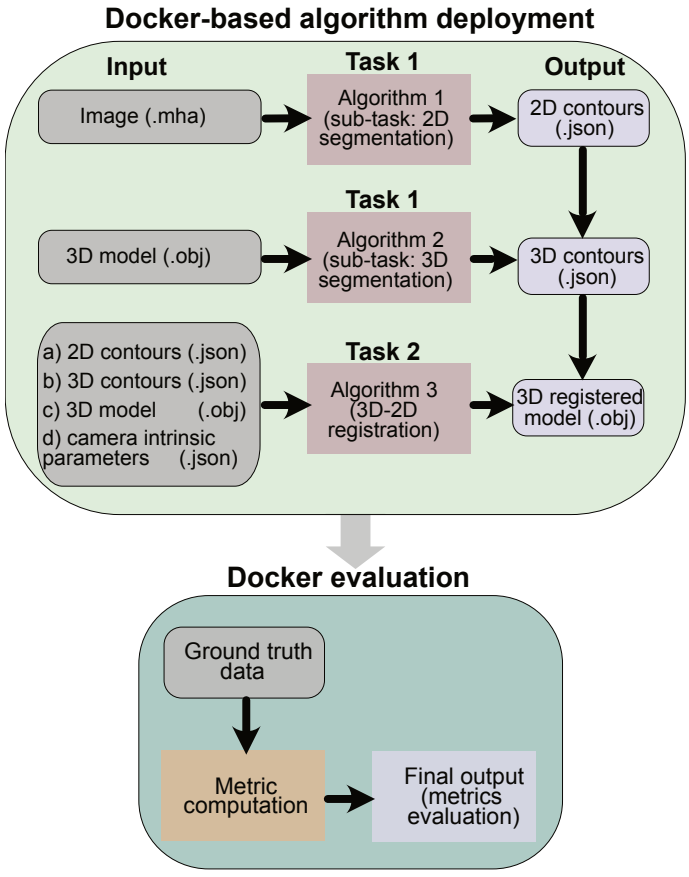

Fig. 4: **Submission procedure of the P2ILF Teamchallenge:** A Docker container system for submission was established on the Grand Challenge platform. Each liver model and corresponding images together with intrinsic camera parameters were provided to the challenge participants. The algorithmic submission required different inputs for the prediction of 2D liver landmarks, 3D liver landmarks, and the use of these landmarks and camera intrinsic parameters for registration of the 3D model to the laparoscopic images. Finally, the outputs from each team's algorithm were evaluated using different metrics (see the section Evaluation Metrics for more details).

Team methods

We describe the methods proposed by each participating team. We explain how every method deals with each of the two tasks, namely the landmark segmentation task and the 3D preoperative to 2D laparoscopic image registration task. At the end of the section, we provide in table 2 a summary of the 2D-3D landmark segmentation strategies proposed for Task 1 and in table 3 the 3D preoperative to 2D laparoscopic image registration strategies proposed for Task 2. The participating teams in the challenge were the *BHL* team from Fudan University (China), the *UCL* team from University College London (United Kingdom), the *GRASP* team from the University of Pennsylvania (United States), the *VOR* team from the University of Science and Technology of China (China), the *NCT* team from the National Center for Tumour Diseases in Dresden (Germany), and the *VIP* team from the Hong Kong University of Science and Technology (China).

Team 1 (BHL team)

The BHL team has proposed an automatic way of segmenting the 2D and 3D landmarks using deep learning methods for the first task and a classical semi-automatic rigid registration approach that uses the segmented landmarks for the second task.

Task 1: a) Segmentation of 2D landmarks

Preprocessing: A Fast Fourier Transform was first applied on the original images. Then, an Inverse Fourier Transform was applied on the high-frequency components to obtain contour-enhanced images (Yang and Soatto, 2020). The team found that this contour enhancement improved segmentation of the *silhouette* and the *falciform ligament* landmarks, but not of the *ridge* landmark. The images were resized to 256×512 pixels for GPU acceleration purposes. The ground truth labels were extended by three pixels using an adjacent pixel strategy.

Data augmentation: Photometric and geometric transformations were applied to the training dataset, namely variations in brightness, contrast, random noise, scaling, cropping, clipping, and rotation.

Algorithm: Two separate ResUnet (Zhang et al., 2018) were used. The first one segmented the *ridge* landmark from the original image, and the second one segmented the *silhouette* and the *falciform ligament* landmarks from the contour-enhanced image. The resulting segmentations were dilated by three pixels.

Loss function: A Dice loss and a cross-entropy loss were used to train each of the ResUNet models. A single L1 loss L_B was introduced at the end to improve the consistency of the two models:

$$L_B = |m_1 - m_2|, \tag{1}$$

where $\{m_1, m_2\}$ are the output maps of the first and second ResUNet, respectively.

Pretraining: No pretraining was done.

Task 1: b) Segmentation of 3D landmarks

Preprocessing: To deal with the class imbalance problem in the mesh data, the groundtruth landmarks were dilated twice using a distance threshold of 20 mm. The vertices in each mesh were then normalised as follows:

$$(x, y, z) = \left(\frac{x_i - x_{mean}}{x_{max} - x_{min}}, \frac{y_i - y_{mean}}{y_{max} - y_{min}}, \frac{z_i - z_{mean}}{z_{max} - z_{min}} \right), \tag{2}$$

where *mean* is the average coordinates of all vertices, *max* and *min* are the maximum and minimum coordinates, respectively.

Data augmentation: The mesh dataset was augmented by applying random rotations and scales (0.75 to 1.25 times) to mimic the liver's size and orientation changes.

Algorithm: A PointNet++ network (Qi et al., 2017) was used to segment the *ridge* and the *falciform ligament* landmarks.

Loss function: A cross-entropy loss function was used to train the PointNet++ network.

Pretraining: No pretraining was done on the PointNet++ network.

Table 2: Summary of the participating teams in **Task 1** of the P2ILF Challenge (2D-3D landmark segmentation)

Team	Algorithm		Loss function		Preprocessing		Data aug.		Pretraining	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
BHL	ResUnet	PointNet++	DSC+CE $+l_1$	CE	FFT, IFT, dilation	Mesh norm.	Yes	R/S	No	No
NCT	nnUNet	MeshCNN	CE+DSC	CE	Dilation	Label merge, Mesh norm.	No	ST	No	No
UCL	UNet++	PointNet++	DSC +Hfd	Hfd +NLL	No	No	ST	SP	ST	No
VIP	Att. UNet	No	CE + IoU^0	No	Resizing	No	Yes	No	No	No
VOR	Various	GCN	CE	CE	No	Mesh norm.	No	VM	No	No

Various: UNet + YOLOv5 + DINO + DeepLabV3; Att.: Attention; CE: Cross-Entropy; FFT: Fast Fourier Transform; IFT: Inverse Fourier Transform; Hfd: Hausdorff distance; DSC: Dice similarity coefficient; IoU: Intersection over union; norm.: normalisation; S: scaling, R: rotation; ST: synthetic data; ST: synthetic data; SP: Spectral augmentation; VM: Vertex masking

Table 3: Summary of the participating teams in **Task 2** of the P2ILF Challenge (3D preoperative to 2D laparoscopic image registration)

Team	Algorithm	Initialisation	Registration constraints	Loss function	Type
BHL	Iterative P*n*P	None	3D-2D Ridge	2D reprojection error	Rigid
GRASP	Iterative Diff. Render.	Average pose estimation	2D Silhouette	2D reprojection error	Rigid
NCT	Iterative Diff. Render.	Constrained random initialisation	3D-2D Ridge + Ligament	2D reprojection error	Rigid
UCL	Iterative Diff. Render.	Fixed initialisation	3D-2D Ridge + Ligament + 2D Silhouette	2D image similarity + 2D Chamfer loss	Rigid
VOR	Multi-staged Spatial Transformers + Diff. Render.	None	Visible liver surface	2D image simialrity + 3D shape-based regularisation	Rigid

P*n*P: Perspective-n-Point; Diff. Render.: Differential rendering

Task 2: 3D preoperative to 2D laparoscopic image registration
Initialisation: The team did a random initialisation.
Algorithm: The team used the iterative P*n*P algorithm from the OpenCV library, along with the intrinsic camera parameters. The obtained rigid transformation, described by a rotation $R \in SO(3)$ and a translation $t \in \mathbb{R}^3$, was used to register the 3D model into the image.
Registration constraints: The 2D and 3D *ridge* landmarks were used as constraints. The segmented landmarks were manually sampled to have the same number of points. These point correspondences served as input to the P*n*P algorithm.
Loss function: The iterative P*n*P algorithm uses reprojection errors to estimate the transformation parameters.
GPU usage: The team used an NVIDIA GeForce RTX 3080 for training their model in Task 1 (21 hours for training time on the 2D segmentation sub-task and 19 hours on the 3D segmentation sub-task), while no GPU was used for Task 2.

Team 2 (UCL team)

Task 1: a) Segmentation of 2D landmarks
Preprocessing: No preprocessing of the training dataset was done.
Data augmentation: A set of synthetic liver images was generated using Unity and Blender to complement the provided training dataset. 100,000 images were generated using 3D liver models and textures purchased from the Unity Asset Store (Unity Technologies, 2023), and textures taken from freely available sources. For every image, random values were uniformly sampled for texture, camera position, lighting effects, motion blur and lens distortion. For each of the 9 patients in the training set, 1000 extra images were simulated in Blender using the patient-specific liver models.
Algorithm: A UNet++ (Zhou et al., 2018) was used to segment the 2D anatomical landmarks. After pretraining, the model was fine-tuned using the patient data for 10 epochs with an ADAM optimiser and an adaptive learning rate starting from 10^{-6}.
Loss function: A combination of Dice loss and Hausdorff distance was used for training.
Pretraining: The UNet++ was pretrained using the synthetic data for 10 epochs.

Task 1: b) Segmentation of 3D landmarks
Preprocessing: No preprocessing of the training dataset was done.
Data augmentation: Spectral augmentation (Foti et al., 2020) was performed to produce a broader collection of 3D models. In addition to the 9 preoperative 3D models provided in the challenge, the team also used the phantom model from Espinel et al. (2022). For 9 of these models, 199 augmentations were generated, giving a total of 1800 models for training. The remaining patient was also augmented with 199 extra models, giving a total of 200 models for validation.
Algorithm: Geometric deep learning was used to segment the 3D landmarks through PyTorch Geometric and PointNet++.

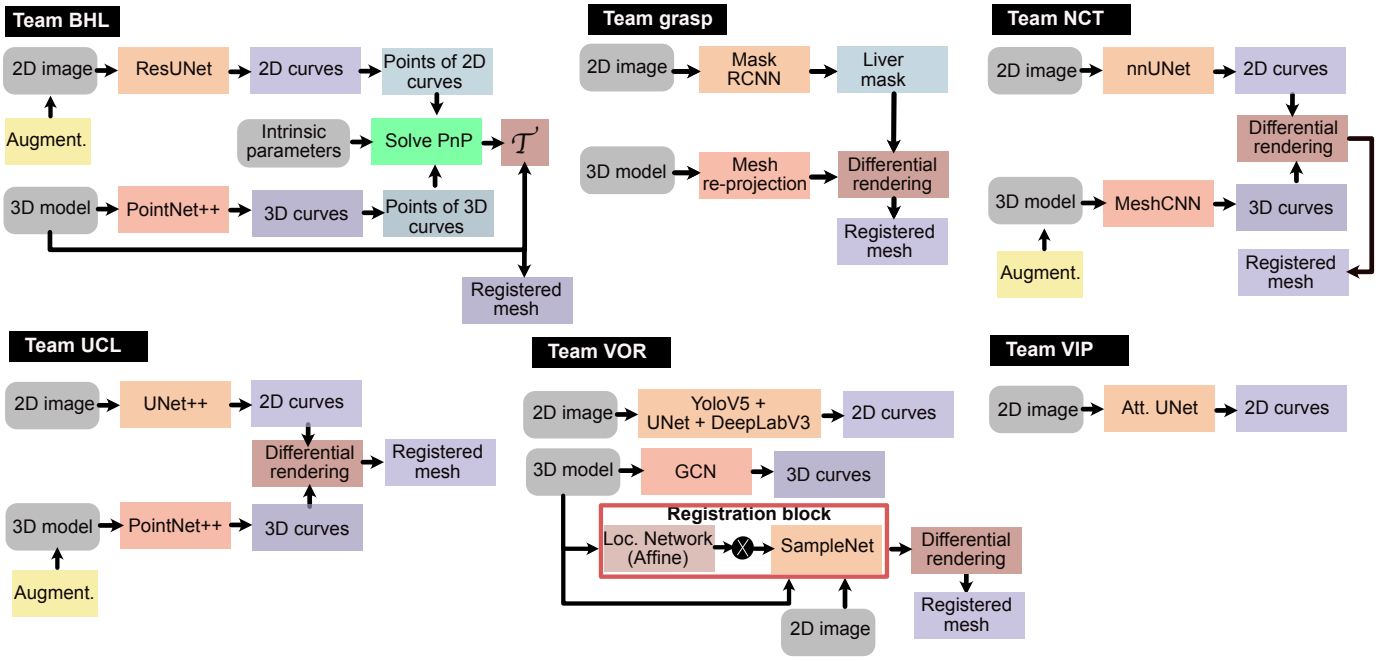

Fig. 5: **General pipeline of the six team methods. Team BHL:** The input 2D image and 3D model are first processed and augmented. Two ResUNets are used to segment the 2D landmarks in the images, and one PointNet++ is used to segment the 3D landmarks in the preoperative 3D model. To perform 3D preoperative mesh to 2D laparoscopic image registration, the correspondences are fed to the P*n*P algorithm and a transformation matrix is obtained. **Team GRASP:** Mask-RCNN is used to generate a 2D mask of the liver, which is then used to perform 3D preoperative mesh to 2D laparoscopic image registration by minimising a *silhouette* reprojection error through differentiable rendering. **Team NCT:** nnUNet and MeshCNN are used to segment the 2D and 3D landmarks, respectively. Differential rendering is then used to perform preoperative 3D mesh to 2D laparoscopic image registration by minimising a reprojection error of the previously segmented landmarks. **Team UCL:** UNet++ is used to segment the 2D landmarks, while PointNet++ is used to segment the 3D landmarks. This team also used differential rendering to perform image registration. **Team VOR:** The 2D case is treated as a pixel segmentation task and the 3D case as a vertex classification task. Differentiable rendering is then used to perform 3D preoperative to 2D laparoscopic image registration by generating 2D images from the affine transformations computed by the localisation networks. The shape regularisation terms provide extra supervision to avoid undesired mesh deformations. **Team VIP:** The team only participated in task 1. Attention UNet was used for the pixel segmentation task of the anatomical liver landmarks in the laparoscopic images.

Training was conducted over 1000 epochs with an ADAM optimiser, using a learning rate of 10^{-3}.

Loss function: A global loss combining Hausdorff distance and Negative Loss Likelihood (NLL) was used for training.

Pretraining: No pretraining was done on the PointNet++ network.

Task 2: 3D preoperative to 2D laparoscopic image registration

Initialisation: The team used a fixed initialisation for which the position of the liver model was initialised with $R = [0, 0, 0]$ and $T = [0, 0, 500]$.

Algorithm: The team proposes a differential rendering pipeline using PyTorch3D. The pipeline iteratively renders the *silhouette* of the preoperative liver model M. The position of the liver model is initialised with a rotation $R = I_3$, where I_3 is a 3×3 identity matrix, and a translation $t = [0, 0, 500]$. An initial registration process is carried out over 100 iterations, where every iteration is performed in five steps. First, the 3D liver model is rendered based on the current R and t. Then, the *silhouette* is extracted by sweeping every column of the image and setting the first non-zero pixel to one while making the other pixels to zero. Second, an image loss is computed between the rendered 3D *silhouette* and the 2D *silhouette* landmark segmented with the method proposed in Task 1. Third, all the points of the 3D *ridge* and *falciform ligament* landmarks segmented in the first task are projected in 2D. Fourth, a Chamfer loss is computed between the projected 3D landmarks and the corresponding 2D landmarks. Fifth, the image and Chamfer losses are backpropagated through the network to update R and t. After the first 100 iterations, a rough initial alignment is achieved, which is used to identify point correspondences between the 3D and 2D landmarks extracted from the first task. After this, another 25 iterations are carried out, with the difference that only the 3D point correspondences found in the initial alignment are used.

Registration constraints: The 2D and 3D *ridge*, *falciform ligament*, and *silhouette* landmarks were used to constrain registration.

Loss function: An image similarity loss and a Chamfer loss were used to estimate the transformation parameters.

Team 3 (GRASP team)

Task 2: 3D preoperative to 2D laparoscopic image registration

Initialisation: The optimisation process began by initializing the mesh to a canonical pose of the organ with respect to the camera. The canonical pose was calculated by manually registering meshes to 15 images and taking an average of the ground truth poses.

Algorithm: First, a Mask R-CNN network was used to segment the liver region. Then, A differentiable rendering approach is used to rigidly register the preoperative 3D model to the laparoscopic image. A transformation T registers the preoperative 3D model M using a rotation R and a translation t. The registration

process begins by initializing the mesh to the canonical pose. Using Pytorch3D's differentiable rendering module, for each optimisation step j a *silhouette* image is rendered using the 3D model transformed by T_j. By back-propagating the loss between the rendered *silhouette* and the predicted *silhouette* from Task 1, a new pose T_{j+1} is computed. The 3D model is then registered in the next step using this new pose.

Registration constraints: The silhouettes of the projected 3D liver model and the segmented 2D liver are used to constrain registration.

Loss function: The optimal transformation T^* is computed for every image by minimising a reprojection error:

$$T^* = \arg\min_T E(T, M, S), \tag{3}$$

where $E(T, M, S) = L_\epsilon(D(T(M)) - S)$ is the reprojection error function, λ is a weighting term, L_ϵ is the smooth L_1 loss, D is the differential rendering function (Ravi et al., 2020), and S is the predicted liver mask.

GPU usage: The team used an NVIDIA Tesla P100 for training their model with 20 minutes training time.

Team 4 (VOR team)

Task 1: a) Segmentation of 2D landmarks

Preprocessing: No preprocessing was done on the training dataset.

Data augmentation: No augmentation of the training dataset was done.

Algorithm: The team proposed a multi-staged network for each type of anatomical landmark, incorporating a UNet pretrained on the provided dataset (Ronneberger et al., 2015) to perform an initial segmentation, along with a YOLOv5 (Redmon et al., 2016) as region proposal module. The anatomical segmentation from the UNet is first converted to a box-shaped segmentation mask. This mask is then combined with the results from the YOLOv5 to form a region-of-interest (ROI) from where representative features are learned for the final segmentation. Then, this ROI mask is multiplied with the original RGB image, and the resulting patch is downsampled from 1920×1080 pixels to 960×540 pixels. A DINO transformer (Caron et al., 2021) is used to generate feature representations from the previously generated patches. Lastly, a DeepLabV3 network (Chen et al., 2017) is implemented to segment the final anatomical landmarks.

Loss function: The DeepLabV3 network uses cross-entropy loss to perform semantic segmentation.

Pretraining: The team did not use models pre-trained on other datasets.

Task 1: b) Segmentation of 3D landmarks

Preprocessing: The 3D models were normalised, i.e. converted into unit space, to improve training stability.

Data augmentation: Random vertex masking was used to augment the 3D models.

Algorithm: A Graph Convolutional Network (GCN) (Kipf and Welling, 2016) is used to segment the mesh vertices. The dataset is split on a per-patient basis in 80% for training and 20% for validation purposes. The team verified that all the images of each patient were contained in either the training or the validation set.

Loss function: The GCN network uses a cross-entropy loss to perform semantic segmentation.

Pretraining: No pretraining of the GCN network was done.

Task 2: 3D preoperative to 2D laparoscopic image registration

Initialisation: Random initialisation was used.

Algorithm: Sampling-based localisation networks are used to perform registration. The approach is designed to deal with two main problems. First, correlating the 2D image with the 3D mesh and, second, preserving the mesh topology and volume during registration. The localisation networks are inspired by the Spatial Transformer network (Jaderberg et al., 2015). They learn a parameterised affine transformation T at every stage, which is then applied to the preoperative liver model M. Then, a sampling module projects the visible vertices onto the images and associates the projected vertices with colours. A Soft Rasterizer (Liu et al., 2019) generates an image from the projected vertices in order to compute an image similarity loss. In addition, the team observed that decoupling the transformation prevents the model to generate unsuitable affine transformations.

Registration constraints: The shape of the projected 3D liver model is used as constraint by comparing it to the liver in the laparoscopic image.

Loss function: A 2D image similarity loss is combined with a 3D shape regularisation term for training. The 2D loss exerts the major supervision when learning the optimal transformation T^* for registration. This dimensional reduction inevitably loses the control of 3D shape properties. To compensate the impact of such reduction, two shape-based regularization terms were added: a Laplacian loss that quantifies the smoothness of the local surface around each vertex, and an edge length loss that penalises significant changes in the edge lengths, avoiding undesired deformations in the mesh such as flattening, erosion, or dilation.

GPU usage: The team did not provide any insight in the GPU usage.

Team 5 (NCT team)

Task 1: a) Segmentation of 2D landmarks

Preprocessing: To capture more information during training, the ground-truth labels were dilated by 10 pixels.

Data augmentation: No augmentation of the training dataset was done.

Algorithm: An nnUNet network (Isensee et al., 2021) was used to perform semantic segmentation. Training is performed using a five-fold cross-validation scheme, which results in five sets of network weights. Since each of the networks generate under-segmented results, they are combined as the union of all the predicted *falciform ligament*, *silhouette* and *ridge* landmarks.

Loss function: nnUnet uses a combination of Dice loss and cross-entropy loss for training.

Pretraining: No pretraining was done on the nnUnet network.

Task 1: b) Segmentation of 3D landmarks

Preprocessing: The per-view landmark annotations are merged into a single *ridge* and *falciform ligament* landmark for each of the patients. The 3D models are also downsampled to improve performance.

Data augmentation: The provided dataset was augmented by deforming the 3D models using finite element simulations. A total of 8208 models with the corresponding labels were obtained, which composed the training dataset. The original undeformed models were used as validation dataset.

Algorithm: Two MeshCNN networks (Hanocka et al., 2019) were used to segment the *ridge* and *falciform ligament* landmarks independently. MeshCNN operates directly on the triangular meshes, extracting local edge features to make predictions that are invariant to rotation, translation, and scale of the input data. Therefore, two sets of weights are used to predict the two classes independently. The union of the two predictions completes the final landmark segmentation. Once the edges have been segmented, each vertex is assigned the class of the edges it is a part of. Prior to any operation, MeshCNN normalises edge lengths by their mean and standard deviation in the dataset. For the normalisation step during inference on unseen test samples, the mean and standard deviation of the original patient data is used.

Loss function: In order to tackle the class imbalance problem, the networks were trained using a cross-entropy loss, with the highest weight assigned to the *falciform ligament* class.

Pretraining: No pretraining was done on the MesCNN networks.

Task 2: 3D preoperative to 2D laparoscopic image registration

Initialisation: The initial pose of the liver is selected at a random position in front of the camera around the positive Z-axis. The initial rotation is set up with liver's anterior side facing the camera, and then a random rotation of less than 90° is applied over each axis. The camera is kept fixed at the origin throughout the whole procedure and only the liver is translated and rotated. The liver scale is also kept fixed.

Algorithm: An optimisation scheme based on differentiable rendering is used. The system renders the liver mesh, the segmented *ridge* and *falciform ligament* landmarks using a virtual camera. The rendered 3D landmarks are then compared with the segmented 2D landmarks to obtain a 2D reprojection error. This error is back-propagated and gradients of rotation and translation are calculated. The gradients are then used to update the rotation R and translation t of the preoperative liver model M. Finally, a new render of the registered model M is made. This process is repeated for 150 iterations, resulting in an iterative 3D pose-optimisation scheme using only 2D pixel-level losses. After the first render is done and to speed up convergence, the position of the rendered 3D *falciform ligament* is compared to the position of the segmented 2D ligament. The 3D liver is then translated parallel to the camera plane until the two *falciform ligaments* overlap in the rendered image. This process is repeated for 30 different random pose initialisations to increase robustness against bad initial alignments. To speed up the process, the laparoscopic and the rendered images were scaled to one fifth of their original size.

Registration constraints: The 3D and 2D *ridge* and *falciform ligament* landmarks are used to constrain the optimisation process.

Loss function: A pixel-level mean squared error is measured between the rendered 3D landmarks and the image landmarks.

GPU usage: The team used an NVIDIA V100 for training their model on task 1 sub-task 2D segmentation (18 hours for each fold, 5 folds were trained). For sub-task 3D segmentation the team used one NVIDIA GeForce RTX2080 and trained for 140 hrs 21 mins, while no training was needed for task 2.

Team 6 (VIP team)

Task 1: a) Segmentation of 2D landmarks

Preprocessing: In order to reduce the computation time, images were resized to 272x480 pixels.

Data augmentation: The provided dataset was augmented by applying random flipping, Gaussian noise, Gaussian blur, and light adjustment.

Algorithm: Attention UNet (Oktay et al., 2018) was used to segment the anatomical landmarks. This network integrates Attention Gates (AG) to UNet to reduce false-positive predictions in irrelevant structures. The dataset was randomly split into training and validation sets with a ratio of 4:1. A cross-validation strategy was followed to select the best checkpoint for inference. Images in each mini-batch were randomly sampled from different patients to ensure diversity. Following Oktay et al. (2018), the gating parameters were initialised so that the attention gates pass through feature vectors at all spatial locations. The network was trained from scratch for 50 epochs with an initial learning rate of 10^{-4} and a batch size of 16. The learning rate was then decreased by 0.9 after every 5 epochs.

Loss function: A cross-entropy loss combined with an IoU loss was used to train the Attention UNet.

Pretraining: No pretraining was done on the Attention UNet.

GPU usage: The team used one NVIDIA GeForce RTX 2080 for training their model in task 1 (1 hours training time on 2D segmentation sub-task)

Results

Evaluation metrics

The metrics used to evaluate the tasks vary according to the nature of the problem to be solved. Task 1 uses Precision Dice Coefficient, and Symmetric Distance (François et al., 2020) to assess the predicted 2D landmarks, along with 3D Chamfer Distance to assess the predicted 3D landmarks. Task 2 uses the 2D Hausdorff Distance to measure the accuracy of 3D preoperative to 2D laparoscopic image registration.

Task 1: Metrics for assessing the 2D and 3D landmark segmentation tasks

Precision

We use precision P to measure the quality of the predicted 2D landmarks at a pixel level. It corresponds to the number of true positives over the total number of predicted pixels (true

positives and false positives). Precision is a commonly used metric in semantic segmentation to evaluate the quality of the predictions (Taha et al., 2014):

$$P = \frac{|TP|}{|TP| + |FP|}, \tag{4}$$

where TP are the true positives and FP are the false positives.

Dice Coefficient

We use Dice coefficient DSC to measure the similarity between the predicted and the ground-truth landmarks. It corresponds to the intersection of the pixels in the predicted and ground-truth landmarks, over the total number of pixels in both landmarks. Dice coefficient is also a commonly used metric in semantic segmentation to evaluate the accuracy of the predictions (Müller et al., 2022):

$$DSC = \frac{2|B_I \cap C_I|}{|B_I| + |C_I|}, \tag{5}$$

where B_I is the set of predicted image landmarks and C_I is the set of ground-truth image landmarks.

Symmetric Distance

We use the Symmetric Distance score proposed by François et al. (2020) to assess the similarity of the predicted and ground-truth landmarks. This score takes five performance criteria into account. First, the ground-truth landmarks should not be missed and there should be no spurious predictions. Second, the predictions should be close to the ground-truth landmarks. Third, each ground-truth landmark should only produce a single prediction. Fourth, the score should be invariant to the image resolution. Fifth, the score should be invariant to the amount of ground-truth landmarks. The score is thus defined as:

$$G = \frac{1}{2|C_I|d_{max}} \left(\sum_{b_I \cap Q} d_S(b_I, c_I \setminus FN) + \sum_{c_I \setminus FN} d_S(c_I, b_I \cap Q) \right) + \frac{|FP|}{|I| - 2|C_I|d_{max}} + \frac{|FN|}{|C_I|}, \tag{6}$$

where G is the symmetric distance score, d_{max} is a tolerance distance that defines if a predicted landmark is spurious or not, $b_I \in B_I$ is a landmark in the set of predicted image landmarks B_I, $c_I \in C_I$ is a landmark in the set of ground-truth image landmarks C_I, Q is the tolerance region around the ground-truth image landmarks defined by d_{max}, FP and FN are the false positive and the false negative predictions, respectively, and $d_S()$ is a symmetric distance function.

3D Chamfer Distance

We measure the similarity between the predicted and ground-truth 3D landmarks by means of a 3D Chamfer Distance. It corresponds to the sum of the squared distances between the nearest neighbour correspondences of the predicted and ground-truth landmarks. The 3D Chamfer Distance d_C is a standard metric used to measure the similarity and completion between two point clouds (Wu et al., 2021):

$$d_C(v, w) = \frac{\sum_v \min_w \|v - w\|^2}{|v|} + \frac{\sum_w \min_v \|v - w\|^2}{|w|}, \tag{7}$$

where $v \in b_M$ are the points in a predicted model landmark b_M, and $w \in c_M$ are the points in the corresponding ground-truth model landmark c_M. $|.|$ denotes the cardinality of a set.

We use the average Chamfer distance F between the predicted and ground-truth landmarks as the evaluation metric:

$$F = \frac{1}{|B_M|} \sum_{b_M} d_C(b_M, c_M), \tag{8}$$

where B_M is the set of predicted model landmarks.

Task 2: Metric for assessing the registration task

We measure the accuracy of the 3D preoperative mesh to 2D laparoscopic image registration done by the participating teams by computing the 2D Hausdorff Distance between the ground-truth 2D landmarks and the 2D projections of the registered ground-truth 3D landmarks. It corresponds to the greatest of all the distances from a point in a ground-truth 3D landmark projected in 2D, to the closest point in the corresponding ground-truth 2D landmark. This metric has became the standard way to measure the similarity and the distance between two curves (Rueda et al., 2014):

$$d_H(v, w) = \max\{\max_v\{\min_w \|v - w\|\}, \max_w\{\min_v \|v - w\|\}\}, \tag{9}$$

where d_H is the Hausdorff distance. The final 2D Hausdorff distance also representing reprojection error in this case (rpe) is measured for both the *ridge* and the *falciform ligament* landmarks:

$$\text{rpe} = \frac{1}{|C_M|} \sum_{c_M} d_H(\Pi(c_M), c_I), \tag{10}$$

where C_M is the set of ground-truth model landmarks and $\Pi(.)$ is the 3D-2D projection function. It should be noted that, while measuring target registration errors is the standard way to quantify the registration accuracy, obtaining such groundtruth data on patients is highly complex. This is because it requires using non-standard devices in the operating room and a rigorous ethical approval process. Thus, we have chosen to use reprojection error as a metric to evaluate the accuracy of the proposed methods.

Quantitative results

Performance comparison for the 2D and 3D segmentation task

For the 2D segmentation step, the quantitative results for the precision and the Dice coefficient (DSC) scores are presented in Table 4. Images 4_21 and 4_22 do not have a visible *falciform ligament* landmark. Therefore, the precision and DSC scores are marked as *NA* (not available) for these two images and the average scores for the *falciform ligament* are computed over 14 images instead of 16. BHL team has the highest overall mean

Table 4: **Evaluation of 2D segmentation of landmarks using region-based metrics:** Precision and dice coefficient scores (DSC) are provided for 16 images from the 2 test cases (patients 4 and 11). Each evaluation metric includes values for the *ridge*, the *falciform ligament*, and the *silhouette* landmarks. The higher the precision and DSC values the better. The best results are in bold, the second best are underlined.

Test image	BHL		NCT		UCL		VIP		VOR	
	Precision	DSC	Precision	DSC	Precision	DSC	Precision	DSC	Precision	DSC
4_3	0.14/0.44/0.53	0.01/0.42/0.63	0.31/0.51/0.47	0.04/0.05/0.61	0.16/0.22/0.51	0.02/0.3/0.55	0.06/0.32/0.25	0.0/0.4/0.38	0.05/0.1/0.23	0.01/0.15/0.36
4_4	0.41/0.47/0.55	0.02/0.4/0.65	0.22/0.0/0.53	0.03/0.0/0.67	0.13/0.16/0.49	0.01/0.23/0.56	0.27/0.33/0.24	0.02/0.41/0.37	0.24/0.14/0.23	0.03/0.19/0.37
4_7	0.27/0.5/0.61	0.01/0.44/0.72	0.07/0.0/0.45	0.02/0.0/0.6	0.05/0.2/0.49	0.01/0.25/0.54	0.06/0.26/0.22	0.01/0.31/0.35	0.17/0.18/0.21	0.03/0.25/0.34
4_11	0.16/0.59/0.56	0.01/0.54/0.54	0.12/0.0/0.3	0.02/0.0/0.42	0.05/0.56/0.5	0.01/0.54/0.55	0.35/0.37/0.24	0.03/0.43/0.37	0.32/0.08/0.21	0.04/0.14/0.35
4_17	0.0/0.0/0.57	0.0/0.0/0.59	0.43/0.0/0.35	0.06/0.0/0.51	0.02/0.14/0.62	0.01/0.22/0.59	0.37/0.0/0.23	0.05/0.0/0.34	0.17/0.0/0.19	0.02/0.0/0.31
4_20	0.3/0.34/0.29	0.02/0.46/0.36	0.27/0.57/0.33	0.03/0.31/0.43	0.18/0.21/0.25	0.03/0.29/0.28	0.0/0.0/0.23	0.0/0.0/0.36	0.22/0.22/0.21	0.03/0.35/0.35
4_21	0.51/NA/0.47	0.06/NA/0.52	0.27/NA/0.5	0.06/NA/0.57	0.02/NA/0.36	0.01/NA/0.4	0.0/NA/0.25	0.0/NA/0.38	0.43/NA/0.2	0.06/NA/0.32
4_22	0.0/NA/0.47	0.0/NA/0.39	0.17/NA/0.31	0.04/NA/0.23	0.0/NA/0.33	0.0/NA/0.31	0.25/NA/0.22	0.05/NA/0.3	0.14/NA/0.17	0.04/NA/0.26
11_2	0.48/0.41/0.46	0.03/0.56/0.46	0.18/0.42/0.46	0.04/0.53/0.57	0.15/0.34/0.32	0.03/0.47/0.34	0.14/0.24/0.14	0.01/0.38/0.21	0.05/0.16/0.1	0.01/0.28/0.16
11_3	0.27/0.41/0.46	0.01/0.49/0.5	0.47/0.44/0.41	0.07/0.59/0.53	0.21/0.26/0.33	0.04/0.37/0.37	0.1/0.25/0.14	0.01/0.4/0.22	0.18/0.15/0.1	0.02/0.26/0.16
11_4	0.43/0.44/0.42	0.02/0.57/0.44	0.36/0.4/0.43	0.04/0.5/0.55	0.19/0.64/0.33	0.03/0.66/0.33	0.08/0.23/0.16	0.01/0.36/0.24	0.05/0.18/0.1	0.01/0.3/0.17
11_5	0.37/0.39/0.41	0.02/0.49/0.43	0.16/0.45/0.42	0.03/0.55/0.54	0.18/0.7/0.33	0.04/0.72/0.34	0.07/0.26/0.17	0.01/0.39/0.25	0.25/0.18/0.1	0.01/0.3/0.16
11_6	0.22/0.38/0.43	0.01/0.46/0.47	0.14/0.52/0.39	0.02/0.53/0.52	0.16/0.68/0.33	0.05/0.7/0.34	0.02/0.24/0.14	0.0/0.37/0.21	0.0/0.17/0.12	0.0/0.27/0.18
11_7	0.41/0.38/0.47	0.03/0.36/0.51	0.16/0.34/0.48	0.04/0.39/0.62	0.12/0.55/0.35	0.01/0.58/0.37	0.04/0.23/0.18	0.01/0.37/0.27	0.0/0.16/0.12	0.0/0.26/0.2
11_8	0.4/0.4/0.4	0.01/0.39/0.41	0.13/0.33/0.37	0.04/0.43/0.46	0.18/0.66/0.31	0.01/0.68/0.31	0.14/0.23/0.14	0.02/0.36/0.21	0.05/0.17/0.09	0.01/0.27/0.14
11_9	0.0/0.6/0.32	0.0/0.45/0.34	0.12/0.4/0.37	0.03/0.55/0.46	0.01/0.7/0.25	0.0/0.72/0.26	0.0/0.24/0.13	0.0/0.38/0.2	0.0/0.19/0.11	0.0/0.32/0.17
Mean	**0.27**/0.41/**0.46**	0.02/0.43/0.50	0.22/0.31/0.41	**0.04**/0.32/**0.52**	0.11/**0.43**/0.38	0.02/**0.48**/0.40	0.12/0.23/0.19	0.01/0.33/0.29	0.15/0.15/0.16	0.02/0.24/0.25
Total mean	**0.38**	**0.32**	0.31	0.30	0.31	0.30	0.18	0.21	0.15	0.17

Table 5: **Segmentation of 2D landmarks using distance metric:** The symmetric distance score G is provided for 16 images from the 2 test cases (patients 4 and 11). Each evaluation metric includes values for the *ridge*, the *falciform ligament*, and the *silhouette* landmarks. The lower the symmetric distance score, the better. The best results are in bold, and the second best are underlined. Mean values for each landmark $\bar{G}$ and for the combined overall mean for all landmarks $\bar{G}_{rls}$ are also provided.

Test data	BHL	NCT	UCL	VIP	VOR
4_3	0.67/0.5/0.14	0.33/0.87/0.3	0.61/1.0/0.21	0.77/0.51/0.71	0.65/1.0/0.87
4_4	0.56/0.6/0.14	0.65/1.0/0.08	0.62/1.0/0.34	0.57/0.47/0.73	0.43/1.0/0.79
4_7	0.77/0.45/0.08	0.69/0.84/0.35	0.62/1.0/0.22	0.74/0.71/0.9	0.67/1.0/1.0
4_11	0.76/0.37/0.34	0.7/1.0/0.99	0.72/0.28/0.18	0.58/0.44/0.74	0.59/1.0/1.0
4_17	1.0/1.0/0.35	0.4/1.0/0.85	1.0/1.0/0.37	0.44/1.0/0.89	0.68/1.0/1.0
4_20	0.48/0.77/0.39	0.58/0.43/0.26	0.61/1.0/0.4	1.0/1.0/1.0	0.54/0.88/1.0
4_21	0.36/NA/0.34	0.34/NA/0.33	0.81/NA/0.45	1.0/NA/0.75	0.39/NA/1.0
4_22	1.0/NA/0.57	0.4/NA/0.81	1.0/NA/0.53	0.28/NA/1.0	0.29/NA/1.0
11_2	0.5/0.19/0.5	0.43/0.22/0.31	0.53/1.0/0.63	0.59/0.66/1.0	0.76/1.0/1.0
11_3	0.63/0.19/0.41	0.37/0.16/0.24	0.46/1.0/0.49	0.58/0.65/1.0	0.61/1.0/1.0
11_4	0.63/0.13/0.36	0.44/0.27/0.26	0.6/0.07/0.46	0.68/0.73/1.0	0.85/1.0/1.0
11_5	0.59/0.3/0.38	0.53/0.27/0.32	0.55/0.05/0.46	0.7/0.6/1.0	0.79/1/1
11_6	0.79/0.3/0.38	0.54/0.39/0.51	0.61/0.04/0.47	0.81/0.8/1.0	1.0/1.0/1.0
11_7	0.65/0.58/0.27	0.56/0.46/0.13	0.79/0.11/0.32	0.84/1.0/1.0	1.0/1.0/1.0
11_8	0.81/0.56/0.5	0.63/0.36/0.46	0.81/0.05/0.57	0.65/1.0/1.0	0.76/1/1
11_9	0.82/0.53/0.51	0.46/0.26/0.45	1.0/0.06/0.57	1.0/0.79/1.0	1.0/1.0/1.0
$\bar{G}$	0.69/**0.46/0.35**	**0.50**/0.54/0.42	0.71/0.55/0.42	0.70/0.74/0.92	0.69/0.99/0.98
$\bar{G}_{rls}$	0.50	**0.49**	0.56	0.79	0.87

Table 6: **Segmentation of 3D landmarks:** 3D Chamfer distances for the *ridge* 'r' (**ch_r**), and the *falciform ligament* 'l' (**ch_l**) landmarks are provided for 16 images of the 2 test cases (patients 4 and 11). The ground-truth 3D vertex locations are compared with the predicted 3D vertex locations for the *ridge* and the *falciform ligament*. NA refers to the cases where the landmark is not annotated. The best results are in bold, the second best are underlined. Mean values are computed for all the images except for the failed cases shown in red. The overall mean is computed as an average of ch_r and ch_l for each team. All the distances are given in millimeters.

Test data	BHL		NCT		UCL		VOR	
	ch_r	ch_l	ch_r	ch_l	ch_r	ch_l	ch_r	ch_l
4_3	98.02	69.98	20.14	37.95	7.68	14.45	28.86	22.88
4_4	102.83	61.08	17.82	38.96	16.69	10.92	33.41	24.19
4_7	84.99	55.15	24.59	39.3	10.2	11.95	26.23	20.82
4_11	101.65	52.19	19.46	40.27	15.07	17.21	34.12	22.35
4_17	105.7	78.52	20.76	37.77	7.61	16.43	34.21	30.46
4_20	108.05	78.91	20.55	39.68	17.42	11.91	31.62	18.46
4_21	156.74	NA	36.83	NA	13.17	NA	40.39	NA
4_22	148.92	NA	33.8	NA	38.05	NA	35.5	NA
11_2	146.26	57.72	30.16	33.67	37.49	24.24	29.92	36.62
11_3	136.32	63.43	30.63	35.55	45.13	31.49	28.29	33.54
11_4	145.55	63.31	30.97	37.04	52.76	34.68	37.42	33.88
11_5	152.12	57.72	31.61	32.92	15.28	36.89	32.52	35.13
11_6	143.85	57.86	30.77	33.32	14.23	28.07	30.39	37.64
11_7	145.3	58.32	32.15	34.95	70.82	43.51	33.19	37.59
11_8	152.87	53.45	33.3	34.59	68.01	18.17	37.18	33.31
11_9	172.22	60.76	37.8	33.41	13.23	42.62	43.22	35.15
Mean	128.27	62.02	**27.19**	36.38	27.97	**24.47**	32.90	30.14
Overall mean	95.14		31.78		**26.22**		31.52	

precision of 0.38, and the highest mean values for the *ridge* and the *silhouette* landmarks, with 0.27 and 0.46, respectively. They also obtained the second highest mean precision for the *falciform ligament* landmark (0.41). The second best results are for the NCT and UCL teams, with an overall mean precision of 0.31. The NCT team has the second highest scores for the *ridge* and the *silhouette* landmarks, with 0.22 and 0.41, respectively. The UCL team has the highest score for the *ridge* landmark (0.43). For DSC, a similar trend can be observed, with the BHL team having the best overall mean score of 0.32. However, the UCL team obtained the highest score for the *falciform ligament* landmark, with 0.48. The VIP and VOR team performed poorly on both metrics. Contrary to the precision and the DSC, the lower the symmetric distance score, the better. These results are shown in Table 5. The NCT team has marginally the best overall mean symmetric distance score compared to BHL, with 0.49. They have the lowest mean score for the *ridge* landmark of 0.50, and the second lowest mean scores for the *falciform ligament* and the *silhouette* landmarks, with 0.54 and 0.42, respectively. While the BHL team has obtained the lowest scores for the *falciform ligament* and the *silhouette* landmarks, with 0.46 and 0.35, respectively. In general, we can observe that the *falciform ligament* and the *silhouette* landmarks have the best prediction performances, followed by the *ridge* landmark, which has the worst segmentation performance by all teams.

Table 7: **3D preoperative mesh to 2D laparoscopic image registration:** Reprojection errors in pixels are provided for 16 samples from the two test patients. These errors are computed for the *ridge* (rpe_r) and the *falciform ligament* (rpe_l) between the projected 3D ground-truth landmark vertices in the registered model w.r.t. the 2D ground-truth pixel locations. NA refers to the not available cases. Mean values are computed for all registrations except for the failed cases shown in red. The overall mean is computed as an average of rpe_r and rpe_l for each team. The best results are in bold, the second best are underlined.

Test	**BHL**		**NCT**		**UCL**		**VOR**		**GRASP**	
#1	**rpe_r**	**rpe_l**	**rpe_r**	**rpe_l**	**rpe_r**	**rpe_l**	**rpe_r**	**rpe_l**	**rpe_r**	**rpe_l**
4_3	515.89	702.65	401.36	257.95	525.82	708.6	936.33	499.04	446.93	661.98
4_4	744.36	1050.4	494.53	368.75	732.54	466.69	1035.62	567.85	521.31	762.17
4_7	431.06	398.92	115.73	170.76	656.34	366.39	869.04	480.69	474.44	558.58
4_11	857.2	901.19	360.19	329.4	340.81	479.25	979.86	669.47	443.22	682.49
4_17	500.09	3182.76	323.6	458.22	250.71	442.19	1142.3	853.63	405.36	444.67
4_20	664.21	946.64	183.58	393.21	707.26	300.48	992.22	762	495.07	652.99
4_21	448.63	NA	159.3	NA	799.6	NA	976.17	NA	451.3	NA
4_22	465.4	NA	212.36	NA	604.88	NA	965.59	NA	504.12	NA
11_2	839.91	2019.02	1008.61	356.36	910.76	473.38	1293.85	1194.77	873.09	452.32
11_3	520.78	3115.51	842.67	177.02	613.39	492.87	1253.12	1170.4	794.59	473.93
11_4	508	659.02	720.35	185.74	974.13	697.18	1278.17	574.81	792.05	542.79
11_5	403	688.8	788.44	311.52	825.24	669.36	1281.08	583.92	767.99	608.84
11_6	509.56	710.12	807.11	543.89	936.63	662.15	1283.26	531.75	1107.94	463.24
11_7	489.7	670.1	360.03	408.01	1058.9	816.27	1248.58	511.68	1291.6	1587.21
11_8	388.29	522.38	329.8	237.32	1174.43	824.16	1279.16	472.45	781.51	627.21
11_9	247.95	369.42	361.25	270.65	925.53	679.32	1250.07	746.84	753.88	643.11
Mean	533.38	1138.35	466.80	319.20	752.31	577.02	1129.02	687.09	681.52	654.39
Overall mean	835.86		**393**		<u>664.66</u>		908.05		667.95	

For the 3D segmentation sub-task, the quantitative results are presented in Table 6. The UCL team had the best overall scores, with the lowest distance for the *falciform ligament* landmark at 24.47 mm and the second lowest distance for the *ridge* landmark at 27.97 mm. The NCT had the lowest distance for the *ridge* landmark at 27.19 mm. The VOR team had the second lowest distance for the *falciform ligament* landmark at 30.14 mm.

Performance comparison for the 3D preoperative mesh to 2D laparoscopic image registration task

The quantitative results for the five teams participating in the 3D preoperative to 2D laparoscopic image registration task are presented in Table 7. Reprojection errors are computed for the *ridge* and the *falciform ligament* landmarks. We analyse the mean reprojection errors of both landmarks (where available) as doing it separately does not provide enough information on the accuracy of the registered models. It is worth noting that the camera calibration of the 2 test patients have mean reprojection errors of 0.36 pixels for patient 4 and 0.64 pixels for patient 11. Given the low calibration errors, they are not taken into account to evaluate the registration performance of the teams. Because images 4_21 and 4_22 do not have a visible *falciform ligament*, the corresponding reprojection errors are marked as *NA* and the average scores for the *falciform ligament* are computed over 14 images instead of 16. The registration method of team NCT has the best mean reprojection error with 393 pixels. The team UCL has the second best mean error, with 664.66 pixels. Following closely, the GRASP team has the third best mean error with 667.95 pixels. The VIP team did not participate in this task.

Run time on test data

All the methods except the one from the GRASP team were evaluated using the provided Docker container on an NVIDIA GeForce 1080Ti 11GB GPU. It was observed that NCT had the longest runtime with nearly 197 seconds, divided in 54 seconds for Task 1 a), 11 seconds for Task 1 b), and 132 seconds for Task 2. All other methods took less than a minute, with the UCL team having the lowest time with 16 seconds and the BHL team the second lowest with 25 seconds. Due to the compilation difficulties found for the GRASP team, we used a Google Colab T4 GPU whose overall average registration time was 26 seconds.

Qualitative results

We present a visual representation of the results obtained by the teams in both tasks. For Task 1, we show a side-by-side comparison of the predicted and the ground-truth 2D and 3D landmarks. For Task 2, we overlay the registered 3D models on top of the laparoscopic images of the two test patients.

Segmentation of 2D landmarks

Figure 6 shows the ground-truth and predicted 2D landmarks for three images of the two test patients. The images correspond, from the left most column to the right most column, to the cases 4_7, 4_11, 4_17, 11_3, 11_6, and 11_7 of Table 4. In general, all the teams were able to segment the *ridge*, the *falciform ligament*, and the *silhouette* landmarks in the laparoscopic

images. Visually speaking, the quality of the predictions correspond to the scores reported in Table 4, with the BHL and NCT teams having less spurious predictions compared to the other teams. The *silhouette* landmark is the one with the best predictions across all the teams, with more continuous curves and less missing parts. The *falciform ligament* landmark also has good results with continuous curves and low spurious responses. The *ridge* landmark is the most challenging case, with lots of missing parts and a considerable amount of spurious predictions.

Segmentation of 3D landmarks

Figure 7 shows the ground-truth and predicted 3D landmarks for the same set of images presented in figure 4. The landmarks shown correspond to the *ridge* and the *falciform ligament*. The BHL team was not able to clearly segment the landmarks, segmenting vertices that are far from the ground-truth locations and covering large areas of the liver surface. The NCT team was able to segment the *ridge* landmarks successfully, while the *falciform ligament* landmarks present some spurious responses. The team has segmented the landmarks in the whole 3D model, rather than in a per-image basis, which was the original goal of the task. The UCL team has segmented the *ridge* landmarks successfully for patient 4, with some spurious responses for the *falciform ligament* landmarks. For patient 11, the *ridge* landmarks are not consistent and the *falciform ligament* landmarks are not clearly defined. Although the team segments the landmarks in the whole 3D model and not on a per-view basis, their method gives different responses when run multiple times on the same model. The VOR team was not able to successfully segment the *ridge* landmarks, while the *falciform ligament* landmarks are far from the ground-truth ones and present some spurious responses. The team has also segmented the landmarks in the whole liver, having the same responses at every running instance.

3D preoperative mesh to 2D laparoscopic image registration

Figure 8 shows a fusion of the registered 3D models with the laparoscopic images of figure 6. Matching the results of table 7, the NCT team has the best visual results, with the registered models having a similar pose to the intraoperative livers. However, the models do not exactly fit the boundaries of the intraoperative livers, which means that using them for AR purposes would be inaccurate. The rest of the methods did not provide visually successful results, with the registered models having different poses or being far from the intraoperative livers. Results for the VOR team are not shown due to their registered models falling out of the laparoscopic images.

Comparison of AR images

From the registration results of Task 2, AR images are generated using inner structures like tumours and veins, as shown in Figure 9. The AR images were only generated for NCT as it was the only team that obtained results that were close to reality. Their images are qualitatively compared to an ICP rigid registration method and the baseline method from Koo et al. (2017) that reports a TRE of less than 10 mm. In the left image of patient 4, the rigid registration method shows the left tumour shifted towards the left compared to the baseline method, while the NCT method shows both tumours inside the field of view and closer to each other compared to the baseline method. For the NCT case, the difference is due to the registered liver being shifted to the right of the real liver, and slightly rotated towards the left. In the middle image of patient 4, the baseline method shows that the left tumour is at the border of the liver, and the right tumour is at the left of the vein. For the rigid case, the left tumour is outside of the liver's parenchyma. For the NCT team, the left tumour is outside of the liver and the right tumour is in front of the vein. This is because the registered liver is closer to the camera and slightly rotated to the right compared to the real liver. In the right image of patient 4, the baseline method shows that the left tumour is near the border of the liver and the right tumour is above the vein. For the rigid case, the tumour in the middle is slightly shifted to the right and the vein is shifted upwards compared to the baseline method. For the NCT team, the left tumour is right behind the right tumour and the vein is rotated towards the left. This is because the registered liver is rotated towards the left compared to the real liver. For the 3 images of patient 11, both the baseline and the rigid methods show the tumour and the vein approximately at the same locations. In the left image of patient 11, the baseline method shows that the tumour is in front of the vein. For the NCT team, the tumour looks more extensive, and the vein has a slightly different pose. This is because the registered liver is closer to the camera and slightly rotated upwards compared to the real liver. In the middle image of patient 11, the baseline method shows the tumour being deformed somewhat towards the bottom, due to the ultrasound probe pushing the liver downwards. For the NCT team, the tumour is at the top-left of the vein, while the vein is closer to the bottom border of the liver. This is because the registered liver is rotated upwards compared to the real liver. In the right image of patient 11, the tumour is more deformed than the previous two images, as the probe continues to push the liver downwards. For the NCT team, the tumour and the vein are parallel to each other, with the the tumour being located more to the left compared to the baseline method. This is because the registered liver is rotated slightly to the left compared to the real liver. These results confirm that the NCT method does not follow the movements of the camera and the liver consistently and that, even if the camera is fixed and the liver remains stable, their method can produce different registration solutions. For the rigid method, even if the internal structures seem to be close to the ones shown by the baseline method despite the lack of deformation, their little displacements in the AR image usually mean real world displacements of several millimeters, translating into a wrong guidance to the surgeon.

Ranking

We conducted an aggregate and rank strategy for 2D and 3D landmarks separately, which was then combined based on the ranking consensus across Task1 (Wiesenfarth et al., 2021). It can be observed in Figure 10 that even though team BHL ranked first in the 2D landmark (higher is better in this case), the mean performance is close to that of team NCT, with team UCL only marginally lower. However, for the 3D landmark segmentation,

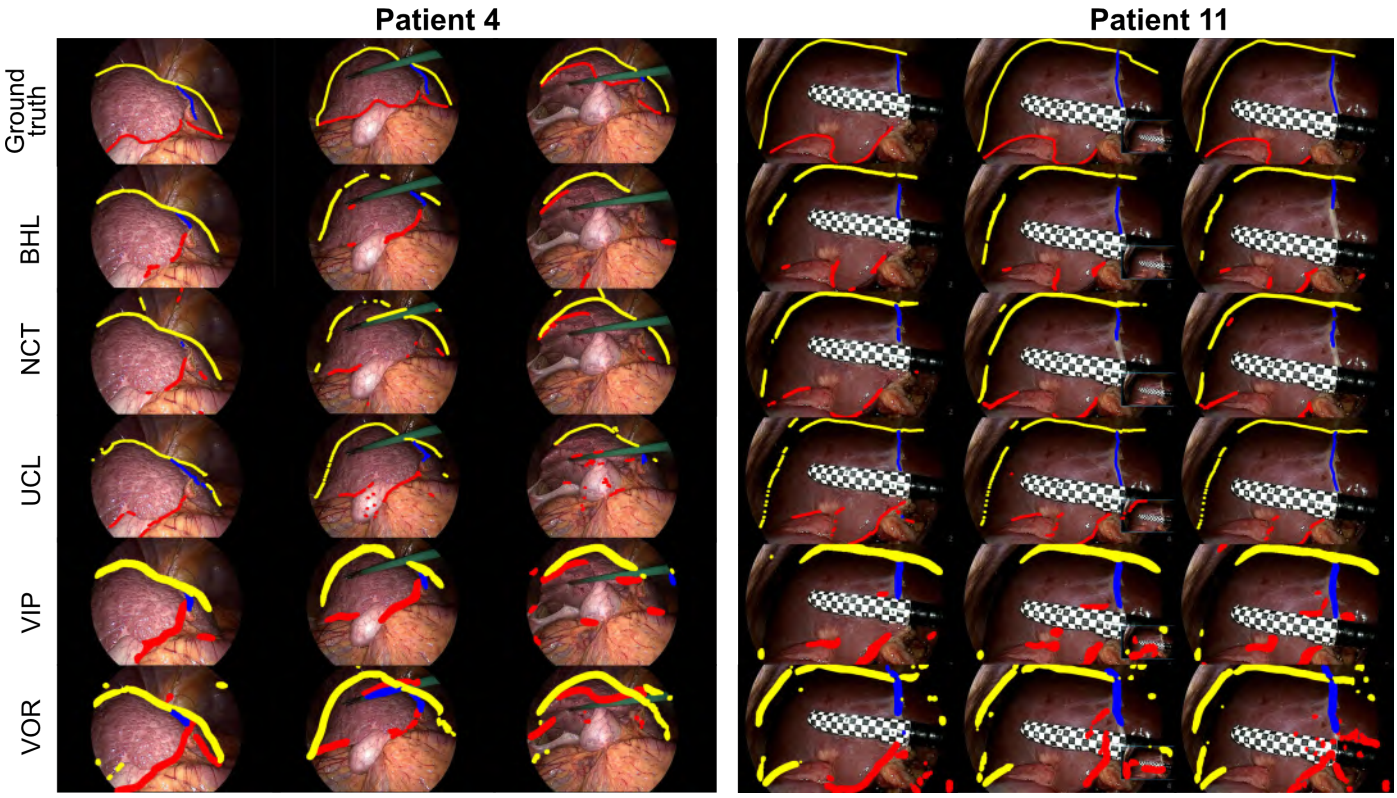

Fig. 6: **Qualitative results of the 2D landmark segmentation task:** The ground-truth (GT) landmarks for the two test patients are shown in the first row, while the teams' predictions are shown in the consecutive rows. The *ridge* landmarks are shown in red, the *falciform ligament* landmarks in blue, and the *silhouette* landmarks in yellow.

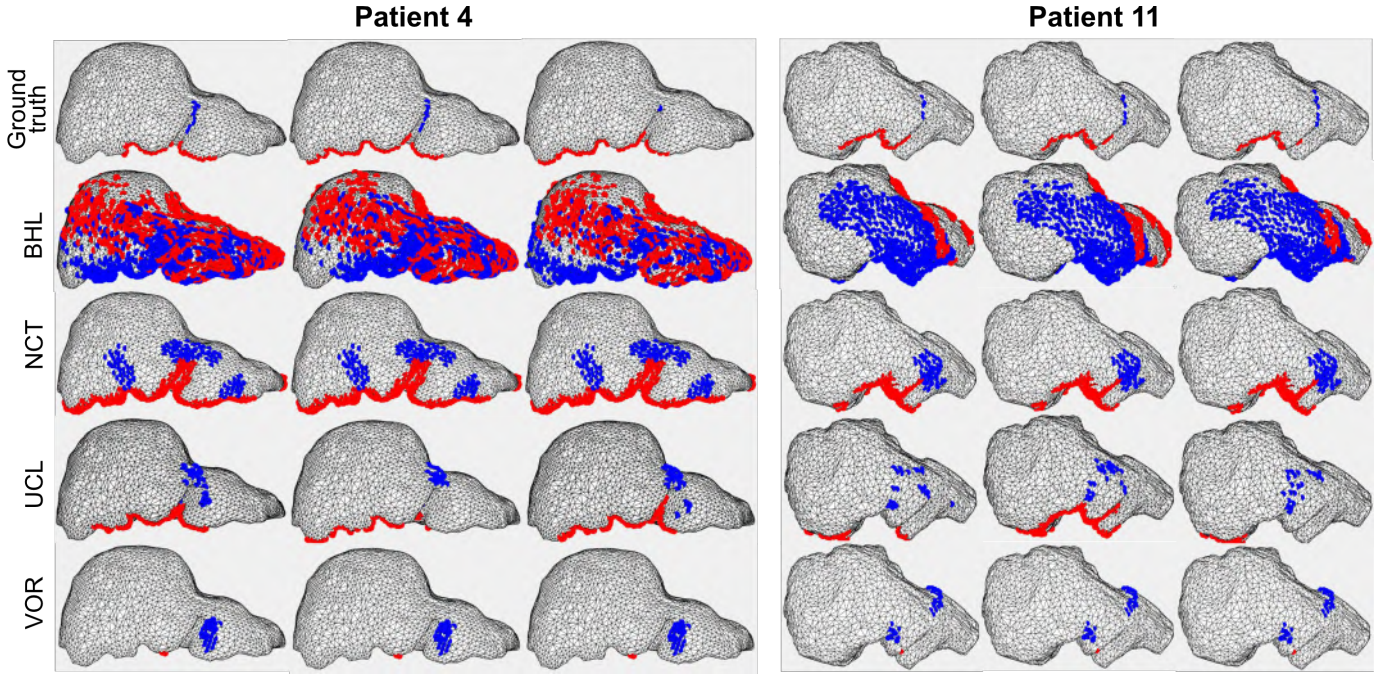

Fig. 7: **Qualitative results of the 3D landmark segmentation task:** The ground-truth (GT) landmarks for the two test patients are shown in the first row, while the teams' predictions are shown in the consecutive rows. The *ridge* landmarks are shown in red and the *falciform ligament* landmarks in blue.

for which lower is better, team UCL outperformed all the other teams, while team BHL had the worst performance. As a result, team UCL's aggregated value in the 2D and 3D landmark segmentation led them to be first in the ranking, with team NCT in second place.

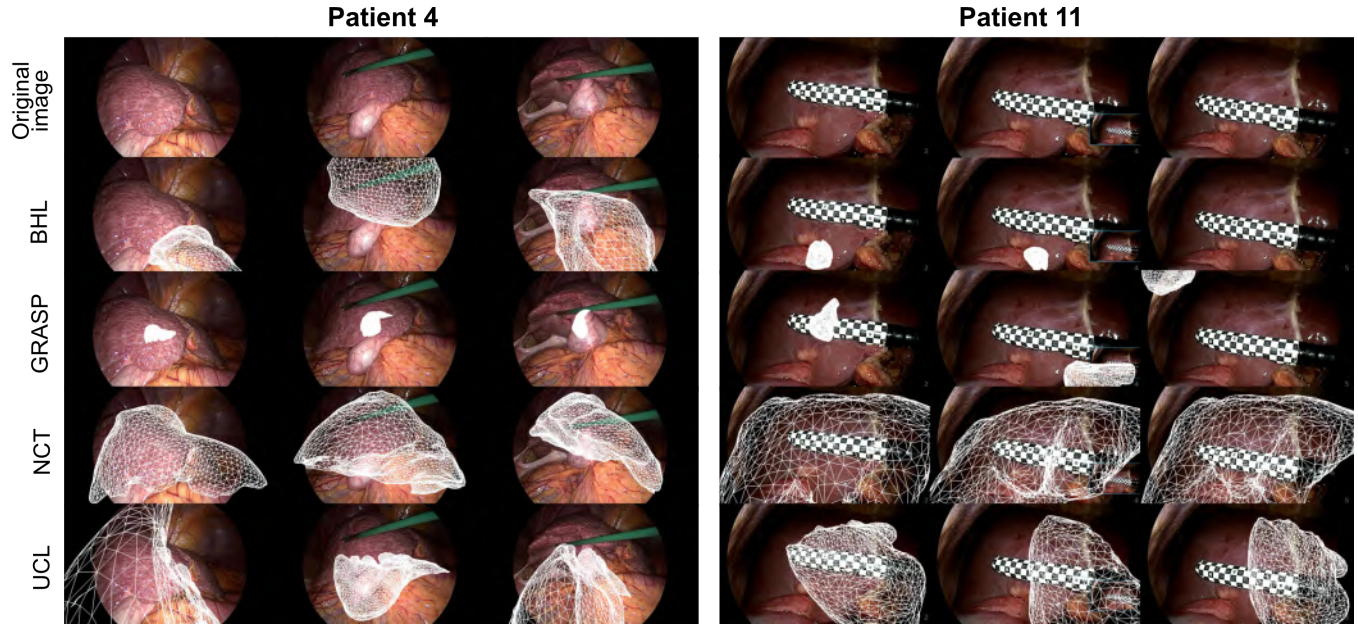

Fig. 8: **Qualitative results of the 3D preoperative mesh to 2D laparoscopic image registration task:** Registration results on some of the images are shown for 4 of the participating teams. The original images are shown in the first row, with the results for BHL, GRASP, NCT, and UCL shown in the consecutive rows. Results for VOR are not shown due to the models being out of the field of view. It can be seen that NCT obtained the best results, with the registered models being close to the liver in the images.

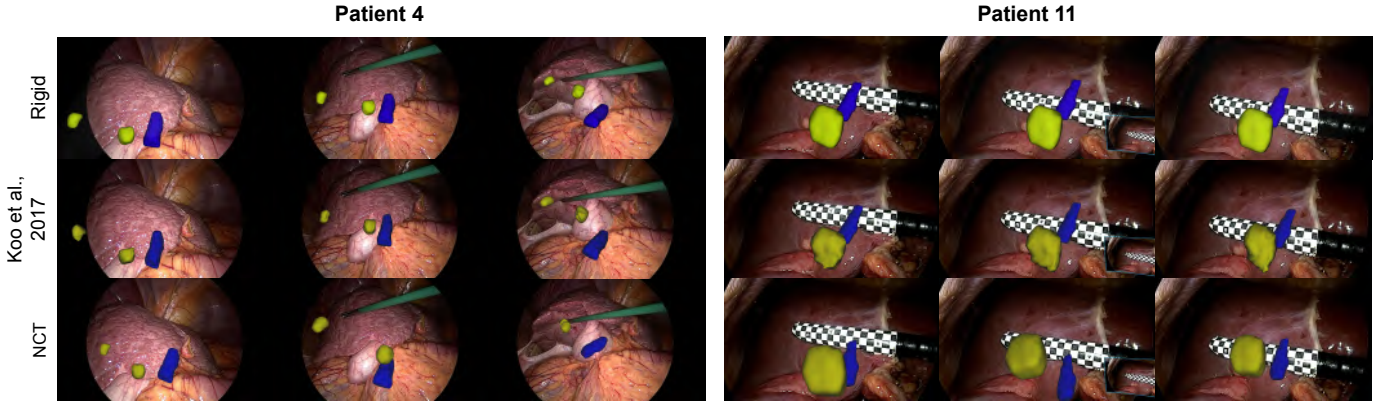

Fig. 9: **Comparison of AR against a baseline method:** AR is generated from the registration results of Task 2 and compared against a baseline method from (Koo et al., 2017). We only compare with the NCT team due to the coherent results they obtained, as shown in Figure 8. Tumours are shown in yellow and inferior vena-cava in blue.

Similarly, on the *Task 2* of the competition and as shown in Figure 11, team NCT outperformed all teams, recording the least registration for most cases with the lowest mean for both the reprojection errors (compared with ridge and with falciform ligament). However, the UCL team ranked third and BHL ranked fourth. These rankings were based on consensus (Wiesenfarth et al., 2021) across reprojection error for ridge and ligament (rpe_r and rpe_l). It is to be noted that in the ranking, we have not taken inference time into account as this challenge is exploratory and requires advancing the registration methods before competing on time requirements.

Discussion

Through this challenge, we aim to release the first comprehensive dataset with carefully annotated anatomical landmarks in both the laparoscopic images and the preoperative 3D models. The 2D landmarks consist of the *silhouette*, the *ridge* and the *falciform ligament* (often persisted as a potential anatomical landmark by surgeons), while the 3D landmarks consist of the *falciform ligament* and the *ridge*. An important limitation of the preoperative to intraoperative registration problem is the validation of techniques, as there is no standard validation strategy. We argue that measuring the reprojection error between 2D and 3D landmarks is a valid strategy, although it carries an ambiguity problem as different registrations can lead to similar reprojection errors. Therefore, it does not fully replace a proper target registration error measurement using reliable 3D

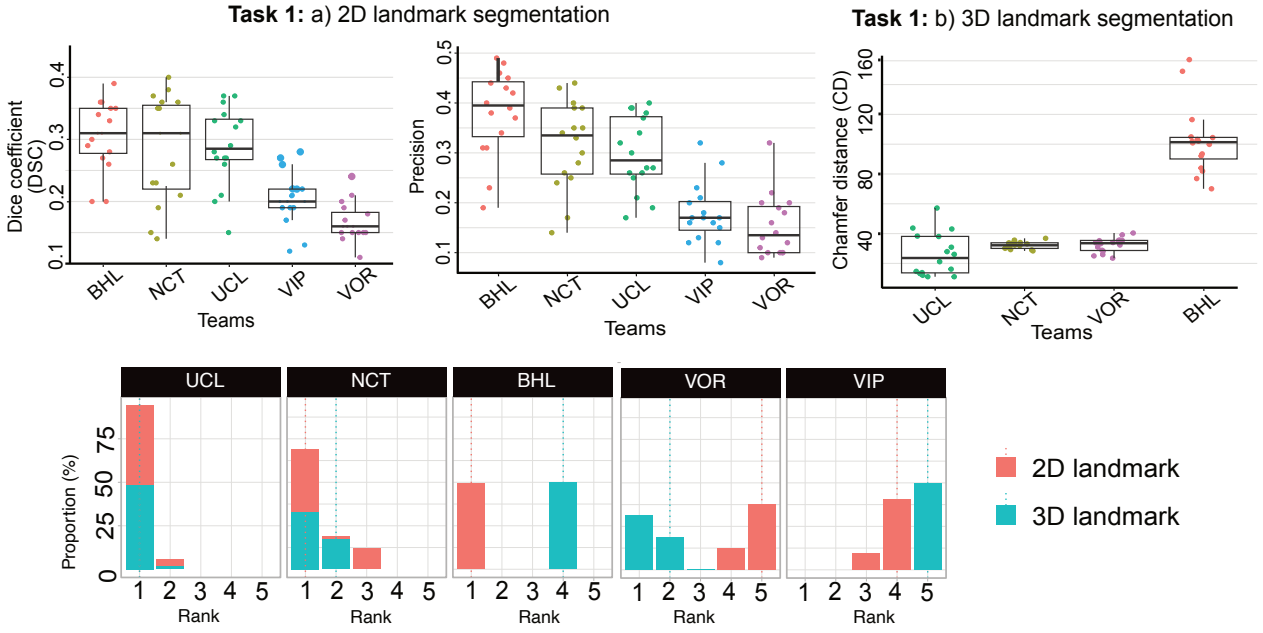

Fig. 10: **Team ranking for *Task 1*:** Dice similarity coefficient and precision are used to perform the ranking of the 2D landmark segmentation task. For the 3D landmark segmentation task, we used the average 3D Chamfer distances between the ridge and falciform ligament (where available). We followed the aggregate and rank strategy for both sub-tasks. At the bottom, we provide a ranking for the proportion of test cases for each team.

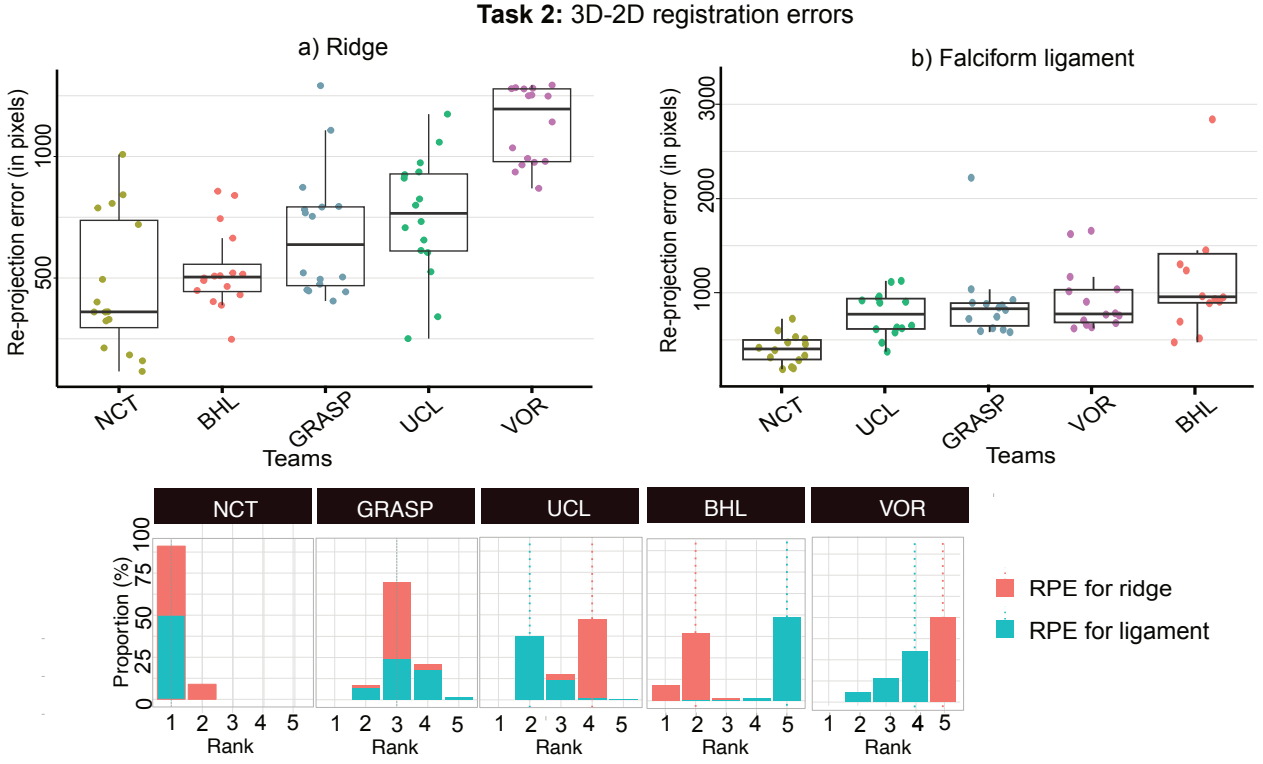

Fig. 11: **Team ranking for *Task 2*:** Reprojection errors (RPE) for 3D-2D registration quantification are used separately - a) with respect to the ridges and b) with respect to the falciform ligament. We followed the aggregate then rank strategy for each. At the bottom, we provide a ranking for the proportion of test cases for each team.

landmarks.

Most participating teams developed methods for both the landmark segmentation and the 3D preoperative mesh to 2D laparoscopic image registration tasks. In terms of 2D landmark segmentation, the majority of teams explored various encoder-decoder UNet-based variants (e.g., Attention UNet, UNet++, nnUNet) either alone or in combination with other models and backbones. In the reports submitted to the challenge organis-

ers, the teams mentioned that using the UNet architecture alone was not enough to have good predictions. For example, the NCT team used 5 fold cross-validation technique using nnUNet and dilation as preprocessing, which encouraged false positives, i.e., penalising precision (second-best score of 0.31 compared to 0.38 from BHL). However, this favored the team in terms of the symmetric distance score with the best score of 0.49. Similarly, the UCL team used UNet++ with heavy data augmentation, generating 100,000 synthetic images for training. The BHL team used two ResUNet, one to predict the *ridge* from the original images, and the other to predict the *falciform ligament* and the *silhouette* from contour-enhanced images. In this way, the team achieved the best precision in all three landmarks, putting them at the top of the ranking for 2D segmentation (Figure 10). *Hypothesis I: Based on the experimental results, segmenting anatomical landmarks in the liver is extremely challenging. Our exploration concluded that using complex model designs or ensemble of models can provide a higher precision, as shown by the BHL team. Using synthetic data for training can improve performance, as shown by the UCL team.*

In the context of the segmentation of 3D anatomical landmarks, all the teams performed a global 3D landmark segmentation instead of a per-view approach, i.e., none of the participating teams utilised the provided 2D laparoscopic view for a given 3D model. Two of the teams (BHL and UCL) utilised PointNet++, the NCT team used MeshCNN, and the VOR team used a graph CNN-based approach. However, since the ratio of the number of annotated vertices to the total number of vertices is very small, teams using off-the-shelf methods without much modification did not succeed in achieving acceptable results (e.g., BHL and VOR). It can be observed that the teams that used simulation techniques to add synthetic meshes for training (team UCL and team NCT) obtained improved 3D landmark segmentation (Table 6). As an aggregate, UCL team ranked first in the 3D segmentation (Figure 10). This is also evident in the qualitative results shown in Figure 7. *Hypothesis II: From the experimental results, it can be concluded that the segmentation of 3D landmarks requires data augmentation to tackle the class imbalance problem. Also, fusing the landmarks from all the views to obtain global ridge and falciform ligament landmarks might help to improve the segmentation performance in 3D models.*

With respect to the 3D preoperative mesh to 2D laparoscopic image registration problem, most of the teams used differentiable rendering as a way to optimise the liver pose. The main difference between them was the registration constraints used during the optimisation. The results from the three teams (BHL, NCT and UCL) on the 2D landmark segmentation and two teams (UCL and NCT) on 3D landmark segmentation are competitive (see Figure 10. However, upon observing the qualitative results for 3D-2D registration (see Figure 8), it can be concluded that only the NCT team's approach provided acceptable registration results. Among these approaches, team NCT obtained the best results both quantitatively (Table 8) and qualitatively (Figure 8), and the only team that had visually satisfactory results. This can be explained by the two distinct approaches they took: 1) The team used an initialisation step, in which the preoperative 3D model is set at a random location in front of the camera, with a rotation such that the liver's anterior side faces the camera and 2) They further constrained their registration using an edge-detection filter in the vertical direction on the projected liver for identifying the silhouette (in addition to the already segmented ridge and falciform ligament), which was not done by other teams. The BHL team was the only one to use a PnP-based approach to perform registration. Apart from the NCT team, the 3D poses obtained by the other teams were far from the liver in the laparoscopic images. Although the reported reprojection errors do not offer a complete overview of the clinical usability of the methods, they do serve as a basis to make an initial assessment, especially for a challenging problem such as LLR. For example, from the quantitative and qualitative results shown in Table 7 and Figures 8 and 9, we can deduce that none of the proposed registration methods will have a successful clinical outcome. Even the best method proposed by team NCT did not show a fully aligned preoperative 3D model and took an average of over 3 minutes. Therefore, conducting a clinical study using the proposed methods would not lead to a successful outcome. *Hypothesis III: From the methods used by the participating teams, it can be deduced that a good initialisation is required to obtain a successful result. This means that the optimisation should start with a pose of the preoperative 3D model close to one of the livers in the image. Otherwise, the methods will converge into a wrong result. Similarly, given the fact that the proposed methods only performed rigid registration, it can be concluded that the methods are not ready for usage in AR, as the deformations between the preoperative and intraoperative stages are not compensated, which also reduces the registration accuracy.*

Conclusion and future directions

The P2ILF challenge is the first challenge that focuses on both the 2D/3D landmark segmentation and the registration problems for AR in laparoscopy. The participating teams understood the importance of the problem and proposed relevant solutions. Although the proposed idea was to segment the 3D landmarks according to the visible 2D landmarks in a laparoscopic image, the teams treated the 2D and 3D segmentation as separate problems. They achieved this by merging the per-image annotated 3D landmarks for all the views of a patient, generating global 3D landmark annotations. Even if it is possible to combine these global 3D landmarks with visible 2D landmarks for registration, using only the per-view visible 3D landmarks may improve registration accuracy. Given the acceptable results for the 2D landmark segmentation and the less accurate 3D segmentation, we can conclude that the 3D segmentation problem is more complex than the 2D segmentation one and requires deeper research. This can be due to the small number of 3D models, 9 for the training set, compared to the number of laparoscopic images, 167 for the training set. Regarding the registration task, using differentiable rendering in combination with the predicted landmarks can provide coherent results, given a good initial pose of the preoperative 3D model. However, the preoperative-to-intraoperative deformations should be taken into account for future approaches to be

used in the operating room. According to these results, a dataset with a larger amount of annotated 2D and 3D data is necessary to improve landmark segmentation. This dataset should be labelled in a way that multiple annotators annotate all the data. Then, an intra- and inter-observer analysis should be done to guarantee the quality of the annotations. Even though methods that used ensemble deep learning techniques and larger iterations performed well, a low inference time is required for clinical adoption. To summarise, a better landmark segmentation combined with preoperative-to-intraoperative deformations should improve the registration of a 3D preoperative mesh into a 2D laparoscopic image, which is important to have an accurate AR.

Data availability. To access the P2ILF challenge dataset, users are requested to create a Synapse account (https://www.synapse.org/) and then use the link to request the dataset as described here: (https://doi.org/10.7303/syn63689257).

Code availability. To help users with the evaluation of the methods, we have shared the evaluation codes used in this manuscript at: https://github.com/sharib-vision/P2ILF.

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

Author contributions. S. Ali, A. Bartoli, Y. Espinel and Y. Jin conceptualised the work, led the challenge and workshop, and prepared the dataset and software. S. Ali and Y. Espinel wrote most of the paper. They did all the analyses with the method contributions from the challenge participants (P. Lue, B. Guttner, X. Zhang, L. Zhang, T. Dowrick, M.J. Clarkson, S. Xiao, Y. Wu, Y. Yang, L. Zhu, D. Sun, L. Li, and M. Pfeiffer) and summarisation assistance from Y. Jin. All authors participated in revising this manuscript, provided input, and agreed to submit it.

References

Adagolodjo, Y., Trivisonne, R., Haouchine, N., Cotin, S., Courtecuisse, H., 2017. Silhouette-based pose estimation for deformable organs application to surgical augmented reality, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 539–544.

Agisoft LLC, 2023. Agisoft Metashape. URL: https://www.agisoft.com/.

Ali, S., Espinel, Y., Jin, Y., Bartoli, A., 2022. Preoperative to Intraoperative Laparoscopy Fusion Challenge (P2ILF) - MICCAI 2022. https://p2ilf.grand-challenge.org/. [Online; accessed 12-August-2022].

Bernhardt, S., Nicolau, S.A., Bartoli, A., Agnus, V., Soler, L., Doignon, C., 2016. Using shading to register an intraoperative ct scan to a laparoscopic image, in: Luo, X., Reichl, T., Reiter, A., Mariottini, G.L. (Eds.), Computer-Assisted and Robotic Endoscopy, pp. 59–68.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660.

Carstens, M., Rinner, F.M., Bodenstedt, S., Jenke, A.C., Weitz, J., Distler, M., Speidel, S., Kolbinger, F.R., 2023. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. Scientific Data 10, 3.

Cheema, M.N., Nazir, A., Sheng, B., Li, P., Qin, J., Kim, J., Feng, D.D., 2019. Image-aligned dynamic liver reconstruction using intra-operative field of views for minimal invasive surgery. IEEE Transactions on Biomedical Engineering 66, 2163–2173.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Espinel, Y., Calvet, L., Botros, K., Buc, E., Tilmant, C., Bartoli, A., 2022. Using multiple images and contours for deformable 3d–2d registration of a preoperative ct in laparoscopic liver surgery. International Journal of Computer Assisted Radiology and Surgery 17, 2211–2219.

Foti, S., Koo, B., Dowrick, T., Ramalhinho, J., Allam, M., Davidson, B., Stoyanov, D., Clarkson, M.J., 2020. Intraoperative liver surface completion with graph convolutional vae, in: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis. Springer, pp. 198–207.

François, T., Calvet, L., Madad Zadeh, S., Saboul, D., Gasparini, S., Samarakoon, P., Bourdel, N., Bartoli, A., 2020. 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.

German Cancer Research Center (DKFZ), 2008. The medical imaging interaction toolkit (mitk). https://www.mitk.org/wiki/The_Medical_Imaging_Interaction_Toolkit_(MITK). [Online; accessed 1-August-2022].

Hanocka, R., Hertz, A., Fish, N., Giryes, R., Fleishman, S., Cohen-Or, D., 2019. Meshcnn: a network with an edge. ACM Transactions on Graphics (TOG) 38, 1–12.

Haouchine, N., Dequidt, J., Berger, M.O., Cotin, S., 2013. Deformation-based augmented reality for hepatic surgery. Studies in Health Technology and Informatics 184, 182–188.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods 18, 203–211.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks. Advances in neural information processing systems 28.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 .

Koo, B., Özgür, E., Le Roy, B., Buc, E., Bartoli, A., 2017. Deformable registration of a preoperative 3d liver volume to a laparoscopy image using contour and shading cues, in: Medical Image Computing and Computer Assisted Intervention, pp. 326–334.

Koo, B., Robu, M.R., Allam, M., Pfeiffer, M., Thompson, S., Gurusamy, K., Davidson, B., Speidel, S., Hawkes, D., Stoyanov, D., Clarkson, M.J., 2022. Automatic, global registration in laparoscopic liver surgery. International Journal of Computer Assisted Radiology and Surgery 17, 167–176.

Labrunie, M., Ribeiro, M., Mourthadhoi, F., Tilmant, C., Le Roy, B., Buc, E., Bartoli, A., 2022. Automatic preoperative 3d model registration in laparoscopic liver resection. International Journal of Computer Assisted Radiology and Surgery 17, 1429–1436.

Liu, S., Li, T., Chen, W., Li, H., 2019. Soft rasterizer: A differentiable renderer for image-based 3d reasoning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7708–7717.

Luo, H., Yin, D., Zhang, S., Xiao, D., He, B., Meng, F., Zhang, Y., Cai, W., He, S., Zhang, W., Hu, Q., Guo, H., Liang, S., Zhou, S., Liu, S., Sun, L., Guo, X., Fang, C., Liu, L., Jia, F., 2020. Augmented reality navigation for liver resection with a stereoscopic laparoscope. Computer Methods and Programs in Biomedicine 187.

Modrzejewski, R., Collins, T., Seeliger, B., Bartoli, A., Hostettler, A., Marescaux, J., 2019. An in vivo porcine dataset and evaluation methodology to measure soft-body laparoscopic liver registration accuracy with an extended algorithm that handles collisions. International Journal of Computer Assisted Radiology and Surgery 14, 1237–1245.

Müller, D., Soto-Rey, I., Kramer, F., 2022. Towards a guideline for evaluation metrics in medical image segmentation. BMC Research Notes 15. doi:10.1186/s13104-022-06096-y.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas, in: Medical Imaging with Deep Learning.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information

processing systems 30.
Radboud University Medical Center, 2023. Grand Challenge. URL: https://www.grand-challenge.org/.
Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G., 2020. Accelerating 3d deep learning with pytorch3d. arXiv preprint arXiv:2007.08501 .
Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.
Robu, M.R., Ramalhinho, J., Thompson, S., Gurusamy, K., Davidson, B., Hawkes, D., Stoyanov, D., Clarkson, M.J., 2018. Global rigid registration of ct to video in laparoscopic liver surgery. International Journal of Computer Assisted Radiology and Surgery 13, 947–956.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.
Rueda, S.L., Sendra, J., Rafael Sendra, J., 2014. Bounding and estimating the hausdorff distance between real space algebraic curves. Computer Aided Geometric Design 31, 182–198.
Soler, L., Nicolau, S., Pessaux, P., Mutter, D., Marescaux, J., 2014. Real-time 3d image reconstruction guidance in liver resection surgery. Hepatobiliary Surgery and Nutrition 3.
Taha, A.A., Hanbury, A., del Toro, O.A.J., 2014. A formal method for selecting evaluation metrics for image segmentation, in: 2014 IEEE International Conference on Image Processing (ICIP), pp. 932–936.
Thompson, S., Totz, J., Song, Y., Johnsen, S., Stoyanov, D., Ourselin, S., Gurusamy, K., Schneider, C., Davidson, B., Hawkes, D., Clarkson, M.J., 2015. Accuracy validation of an image guided laparoscopy system for liver resection, in: Medical Imaging 2015: Image-Guided Procedures, Robotic Interventions, and Modeling, SPIE. pp. 52 – 63.
Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N., 2017. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging 36, 86—97. doi:10.1109/tmi.2016.2593957.
Unity Technologies, 2023. Unity Asset Store. URL: https://assetstore.unity.com/.
Wagner, M., Müller-Stich, B.P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., Lubotsky, D.M., Müller, B., Davitashvili, T., Capek, M., Reinke, A., Reid, C., Yu, T., Vardazaryan, A., Nwoye, C.I., Padoy, N., Liu, X., Lee, E.J., Disch, C., Meine, H., Xia, T., Jia, F., Kondo, S., Reiter, W., Jin, Y., Long, Y., Jiang, M., Dou, Q., Heng, P.A., Twick, I., Kirtac, K., Hosgor, E., Bolmgren, J.L., Stenzel, M., von Siemens, B., Zhao, L., Ge, Z., Sun, H., Xie, D., Guo, M., Liu, D., Kenngott, H.G., Nickel, F., von Frankenberg, M., Mathis-Ullrich, F., Kopp-Schneider, A., Maier-Hein, L., Speidel, S., Bodenstedt, S., 2023. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Medical Image Analysis 86, 102770.
Wiesenfarth, M., Reinke, A., Landman, B., Eisenmann, M., Saiz, L., Cardoso, M., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Scientific Reports 11. doi:10.1038/s41598-021-82017-6.
Wu, T., Pan, L., Zhang, J., WANG, T., Liu, Z., Lin, D., 2021. Balanced chamfer distance as a comprehensive metric for point cloud completion, in: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 29088–29100. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/f3bd5ad57c8389a8a1a541a76be463bf-Paper.pdf.
Yang, Y., Soatto, S., 2020. Fda: Fourier domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095.
Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual u-net. IEEE Geoscience and Remote Sensing Letters 15, 749–753.
Zhong, J.H., Peng, N.F., You, X.M., Ma, L., Xiang, X., Wang, Y.Y., Gong, W.F., Wu, F.X., Xiang, B.D., Li, L.Q., 2017. Tumor stage and primary treatment of hepatocellular carcinoma at a large tertiary hospital in china: A real-world study. Oncotarget 8, 18296–18302.
Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation, in: Deep learning in medical image analysis and multimodal learning for clinical decision support. Springer, pp. 3–11.