ARMADILLO: Augmented Reality Machine-Assisted Detection and Inference in Laparoscopic Liver Operations

Group Report for COMP5530

ABDUL KARIM ABBAS*, JAMES BORGARS*, AODHAN GALLAGHER*, AHMAD NAJMI MOHAMAD SHAHIR*, JIBRAN RAJA*, ABHINAV RAMAKRISHNAN*, and THEODORA VRAIMAKIS*, University of Leeds, United Kingdom
SHARIB ALI† and RAFAEL KUFFNER DOS ANJOS†‡, University of Leeds, United Kingdom

Laparoscopy is an approach to liver surgery which reduces complications and recovery time, and can also harness developments in machine-assisted surgery. In this report, we outline and implement the end-to-end process of performing 3D-2D registration using only a preoperative liver mesh and intraoperative laparoscopy footage, with no human involvement in respect to landmark annotation and alignment. We present novel research in the fields of 2D and 3D landmark segmentation, with best-in-class results for the dataset. We study iterative and deep learning approaches in the area of 3D-2D registration, with silhouette extrapolation implemented for improved results. Finally, we explore hardware implementation of the pipeline and data visualisation techniques using an Augmented Reality headset. Our results include a 30% relative increase in 2D segmentation precision, 36% improvement in 3D segmentation distance, and 31% improvement in reprojection error in registration compared to leading research.

CCS Concepts: • **Human-centered computing** → *Mixed / augmented reality*; *Human computer interaction (HCI)*; • **Computing methodologies** → **Computer vision**; **Image segmentation**; Tracking; *Object detection*; **Supervised learning**; **Neural networks**; *Point-based models*; • **Applied computing** → Life and medical sciences.

Additional Key Words and Phrases: segmentation, registration, deep learning, image-guided intervention, surgical data science, laparoscopy

1 INTRODUCTION

Laparoscopic liver surgery (also known as minimally-invasive liver surgery and keyhole surgery) is a surgical approach which minimises recovery time and the probability of complications [Slakey et al. 2013]. This surgical approach also facilitates developments in the area of machine-assisted surgery due to its use of a camera, such as 3D-2D registration of the liver, where a preoperative 3D mesh of the liver, including anatomical landmarks such as tumours and vessels, can be superimposed onto the liver in real-time during surgery.

In this report, we present an implemented end-to-end pipeline, automating the process of 3D-2D registration of the liver. This includes segmentation models of both preoperative 3D meshes, and intraoperative 2D laparoscopic images, which have both in of themselves warranted novel research currently under review, having outperformed prior research on the same dataset [Ali et al. 2025]. Model outputs are utilised in a registration pipeline that does not require manually annotated data. A visualisation implementation has also been completed to visualise segmentation predictions and create a model navigation environment. Our results have led to substantial improvement in all tasks of the pipeline.

2 BACKGROUND RESEARCH

2.1 2D Segmentation

Ronneberger et al. present the 'U-shaped architecture' for Fully Convolutional Networks (FCNs) in the form of UNet, proving that large datasets were not required for high accuracy in the field of biomedical segmentation [Ronneberger et al. 2015]. The UNet architecture consists of a contracting path, which has pooling layers, an expansive path with up-convolutions, with these two paths connected by a bottleneck and skip connections, achieving increased performance at reduced inference times compared to previous models [Ronneberger et al. 2015].

UNet++ by Zhou et al. builds upon the UNet architecture with a greater number of convolution blocks, dense skip connection pathways, and deep supervision [Zhou et al. 2020]. UNet3+ further develops upon the ideas of UNet++, proposing full-scale skip connections where each convolutional block in the contracting path connects to its opposing and below blocks in the expansive path; the bottleneck and expansive path is supervised by the ground truth and has skip connections to every block further up the path [Huang et al. 2020]. ResUNet is a deep residual UNet-based model, replacing the standard Convolution-ReLU block with residual convolution blocks utilising batch normalisation [Zhang et al. 2018]. Jha et al. propose ResUNet++, modifying the ResUNet architecture for medical image segmentation through the addition of squeeze-excitation blocks to dynamically weight convolutional channels, ASPP for increased context when classifying a pixel, and the introduction of attention for enhanced feature quality [Jha et al. 2019].

Chen et al. propose DeepLabV3+, building on top of the DeepLabV3 architecture with the addition of depth-wise separable convolution to both ASPP and decoder modules, resulting in improved performance, having been tested on non-medical benchmarks [Chen et al. 2018].

Various implementations of UNet have been thoroughly evaluated alongside ResUNet in the Liver Tumour Segmentation (LiTS) benchmark [Bilic et al. 2023], with UNet++ and UNet3+ both demonstrating their outperformance of UNet on the benchmark [Huang et al. 2020; Zhou et al. 2020], highlighting their relevance to the

*Sharing first authorship
†Also with the role of Supervisor.
‡Also with the role of Assessor.

Authors' addresses: Abdul Karim Abbas, sc21aka@leeds.ac.uk; James Borgars, sc20jdb@leeds.ac.uk; Aodhan Gallagher, sc20ag@leeds.ac.uk; Ahmad Najmi Mohamad Shahir, sc21anbm@leeds.ac.uk; Jibran Raja, ed20j3r@leeds.ac.uk; Abhinav Ramakrishnan, sc21a2r@leeds.ac.uk; Theodora Vraimakis, sc20tmv@leeds.ac.uk, University of Leeds, Woodhouse, Leeds, West Yorkshire, United Kingdom, LS2 9BW; Sharib Ali, s.s.ali@leeds.ac.uk; Rafael Kuffner dos Anjos, r.kuffnerdosanjos@leeds.ac.uk, University of Leeds, Woodhouse, Leeds, West Yorkshire, United Kingdom, LS2 9BW.

field of liver surgery. Jha et al. demonstrated that ResUNet++ outperformed both UNet and ResUNet in both the Kvasir-SEG and CVC-ClinicDB datasets, showing its suitability in medical applications [Bernal et al. 2015; Jha et al. 2019, 2020].

Koo et al. demonstrate semantic contour detection of the ridge and silhouette of the liver through the use of Convolutional Neural Networks (CNNs) [Koo et al. 2022], using a CASENet model [Yu et al. 2017] with a ResNet50 encoder [He et al. 2016], pre-trained on the ImageNet dataset [Deng et al. 2009]. Koo et al. augment their dataset through the use of scale, shear, brightness, contrast, rotation, and translation transformations, to increase the generalisation and invariance capabilities of the model [Koo et al. 2022].

As part of the MICCAI 2022 conference, the Preoperative to Intraoperative Laparoscopy Fusion (P2ILF) challenge was hosted, with the goal of automating the end-to-end process of 3D-2D registration in liver laparoscopy surgery [Ali et al. 2025]. This includes the development of methods to perform landmark segmentation on intraoperative laparoscopic footage; all teams published in the challenge decided to use deep learning for this task [Ali et al. 2025]. Differing to the dataset used by Koo et al., the P2ILF dataset has annotations for the ridge, silhouette, and the falciform ligament of the liver [Ali et al. 2025; Koo et al. 2022]. Teams from around the world attempted the P2ILF challenge, covering a range of different approaches in terms of model, loss function, dataset augmentation, and pre-training for the 2D segmentation task [Ali et al. 2025].

Pei et al. introduce the D^{2}GPLAND model, a depth-aware model guided by unified features from an estimated depth map and the original image. The depth map is estimated from the original image using a depth estimation network, which is then put through a frozen Segment Anything Model (SAM) encoder [Kirillov et al. 2023], from which its features are extracted [Pei et al. 2024]. For feature extraction of the original laparoscopic image, a ResNet34 encoder [He et al. 2016] is used [Pei et al. 2024]. D^{2}GPLAND achieves best-in-class results on liver segmentation on their L3D dataset, a collation of laparoscopic images from multiple sources [Pei et al. 2024].

2.2 3D Segmentation

Recent advances in geometric deep learning have enabled segmentation of anatomical surfaces directly on meshes and point clouds. Hanocka et al. introduce MeshCNN, which adapts convolution and pooling operations to irregular triangular meshes by operating directly on edges rather than vertices [Hanocka et al. 2019]. In their approach, edge convolutions aggregate features from neighbouring edges, with learnable filters that are weighted by the angles between adjacent faces. Pooling is performed by collapsing the least relevant edges, allowing the network to retain only important geometric structures throughout the hierarchy. MeshCNN achieves state-of-the-art performance on benchmark segmentation tasks, yet its reliance on handcrafted edge features and sensitivity to mesh resolution can limit scalability on highly detailed anatomical models.

Building on the notion of treating meshes as graphs, Kipf and Welling propose Graph Convolutional Networks (GCNs) for explicit modelling of the mesh topology. In this approach, each vertex becomes a node in a graph and convolution feature propagation is performed by spectral filtering [Kipf and Welling 2017]. This framework captures both fine local geometry and global connectivity, delivering robustness to noise and topological variation. However, dense connections and expanding neighbourhoods leads to high computational and memory costs, making it increasingly difficult to process very high-resolution meshes like those of organs.

An alternative example is to operate on unstructured point sets. Qi et al. present PointNet, which processes each point independently through a shared multi-layer perceptron and achieves permutation invariance via max-pooling [Qi et al. 2017a]. Since max-pooling selects only the maximum activation in each feature channel across the entire point set, it discards the spatial distribution of features from neighbouring points. As a result, fine local structures, such as small curvature variations or intricate surface details, are lost. By concatenating the resulting global feature vector back to per-point features, PointNet delivers competitive segmentation accuracy with remarkable simplicity and efficiency, but its reliance on max-pooling restricts the capture of fine local structures.

To overcome this limitation, Qi et al. extend the architecture hierarchically in a newly revised PointNet++. They introduce set abstraction layers that recursively partition the point cloud via farthest point sampling and ball queries, applying PointNet locally within each region to learn fine-grained descriptors, and then aggregate these across scales [Qi et al. 2017b]. This multi-scale grouping strategy adapts to variable point densities and markedly improves segmentation performance on complex anatomical surfaces by capturing both local detail as well as the global context.

2.3 3D-2D Registration

Registration of the liver has been an area of research for quite some time, as it is known that if successful, clear advances in regards to reducing surgical risk will be made, as these systems could help surgeons identify anatomical structures, particularly in complex interventions [Koo et al. 2022]. Koo et al. explain that the registration process usually occurs in two stages: global alignment, followed by local alignment. Koo et al. mention that automatic approaches to local alignment have been proposed, but rely on good initialisation provided by global alignment, which is usually performed manually to some degree, such as requiring annotation from a clinician during surgery [Koo et al. 2022].

Adagolodjo et al. propose a method that requires manual annotation of the silhouette on the intraoperative image, a process which took around 18 seconds (with an overall intraoperative setup time being around a minute) [Adagolodjo et al. 2017]. This approach, albeit an improvement over manual rigid registration, requires a lot of manual intervention to be considered feasible [Adagolodjo et al. 2017].

The perspective-n-point (PnP) problem is particularly pertinent in the area of registration. PnP, in regards to registration of the liver, is the problem of estimating the laparoscopic camera's pose (position and orientation), given a set of 3D points (i.e. landmarks from the preoperative 3D model), and their respective locations in a 2D image (i.e. landmarks from the intraoperative laparoscopic image). Solutions exist for the PnP problem [Lepetit et al. 2009], but it is usually considered in the perspective of having 3D points and 2D projections of a point concurrently (from the same moment in time), however in this case, the 3D points are from a different time to the 2D points, this combined with the knowledge that the liver is highly deformable during surgery and the factoring in of occlusion, highlights the difficulties in this field. Other approaches in solving this problem can include adding Random Sample Consensus (RANSAC) [Koo et al. 2022] and differential rendering [Ali et al. 2025].

Since abdominal organs can undergo large deformations during surgery [Adagolodjo et al. 2017], it is imperative to be able to model these deformations. Usually approaches utilise the Finite Element Method (FEM) to simulate deformations [Adagolodjo et al. 2017; Labrunie et al. 2024], however As-Rigid-As-Possible (ARAP) surface modelling has also been proposed [Mhiri et al. 2025].

Labrunie et al. propose the Liver Mesh Recovery (LMR) framework [Labrunie et al. 2024]; based off of the Human Mesh Recovery framework [Kanazawa et al. 2018], LMR utilises patient-specific models to perform registration, as opposed to the optimisation-based approach typically adopted by existing solutions, due to the computational expense of these iterative solutions during the procedure [Labrunie et al. 2024]. Labrunie et al. propose training a ResNet model by deforming an annotated preoperative 3D liver model, and then projecting it with its landmarks annotated such that 2D masks can be simulated, with predefined camera intrinsics and liver deformation parameters known [Labrunie et al. 2024]. After training, the model can then receive a segmentation mask from the 2D segmentation model as input, and predict the deformation parameters required to deform the 3D model it was trained on such that it matches the shape captured in the laparoscopic image.

Mhiri et al. propose a solution utilising the same patient-specific training approach taken by Labrunie et al., albeit with modifications [Mhiri et al. 2025]. In place of FEM, ARAP is used as it is less computationally expensive; also, a simple multilayer perceptron (MLP) is trained to perform registration, with results beating LMR on the dataset used [Mhiri et al. 2025].

3 METHOD

3.1 2D Segmentation

3.1.1 Proposed Method. All models ablated for this task (UNet, UNet++, UNet3+, DeepLabV3+, and ResUNet++) are equipped with ResNet34 encoders that have been pre-trained on the ImageNet dataset [Deng et al. 2009; He et al. 2016]. Models are trained on the L3D dataset [Pei et al. 2024], augmented with flip transformations on both the $x-$ and $y-$ axes. The AdamW optimiser was used for training [Loshchilov and Hutter 2019], in conjunction with a learning rate plateau scheduler. The loss functions ablated over consisted of different weightings of cross-entropy loss, Dice loss, Huber loss, and Focal-Tversky loss.

For cross-entropy loss, the following novel logarithmic class weights function was used, with $W(i)$ being the weight for a class i, M is the set of all training mask pixels, and $M(i) \subseteq M$ is the set of mask pixels of class i:

$$W(i) = \max\left[1.0, \log_{10}\left(\frac{|M|}{|M(i)|}\right)\right] \tag{1}$$

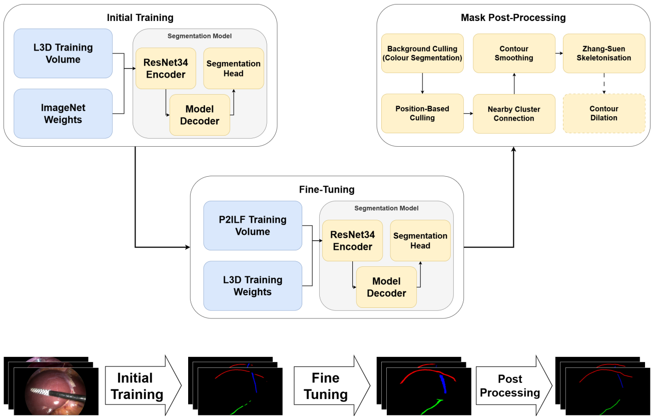

Fig. 1. 2D segmentation task pipeline. Dashed processes are optional.

Once the training ablations have been completed, the models are evaluated using two validation patients from the P2ILF training set [Ali et al. 2025], using precision and Dice similarity coefficient (DSC). Models were then selected to be fine-tuned based on their performance in this evaluation. Fine-tuning is performed on the P2ILF training set (minus the two aforementioned patients which are used as a validation set).

Finally, once the model has been fine-tuned, its predicted masks are then post-processed. This involves culling known false positive pixels (i.e. outside of camera, ligament pixels above/below both the ridge and silhouette), applying the Ramer-Douglas-Peucker algorithm for smoothing landmark contours [Douglas and Peucker 1973; Ramer 1972], and applying the Zhang-Suen algorithm for skeletonisation of contours [Zhang and Suen 1984]. This centre line can then be dilated for evaluation, or sampled into a set of points.

3.1.2 Loss Function. Our proposed composite loss function consisting of weighted cross-entropy loss ($\mathcal{L}_{\text{wCE}}$), Dice loss ($\mathcal{L}_{\text{DL}}$), Huber loss ($\mathcal{L}_{\text{HL}}$), and Focal-Tversky loss ($\mathcal{L}_{\text{FTL}}$), can be represented as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{wCE}} + \beta \cdot \mathcal{L}_{\text{DL}} + \gamma \cdot \mathcal{L}_{\text{HL}} + \delta \cdot \mathcal{L}_{\text{FTL}} \tag{2}$$

In Equation 2, α, β, γ, and δ are the hyperparameters which we intend to ablate to find the optimal solution.

3.2 3D Segmentation

3.2.1 Proposed Method. The PointNet++ architecture used in our approach is designed to process and segment 3D point clouds directly, making the conversion of 3D data into point cloud representations a necessary step in our method. The implementation used is adapted from the repository by Yan [Yan 2019] and is based on the work demonstrated by Qi et al. [Qi et al. 2017b]. The PointNet++ network was trained on the datasets using different loss functions, including a novel midline loss. The optimal hyperparameters and loss functions were selected through an ablation study, detailed in Section 4.2.3. The standard PointNet architecture was also tested [Qi et al. 2017a].

3.2.2 Loss Function. We introduce a novel midline loss function, $\mathcal{L}_{\text{midline}}$, designed to encourage alignment between predicted segmentation points and a central anatomical midline derived from ground-truth labels. This loss improves structural consistency in thin and elongated anatomical regions.

Let $\mathcal{X}$ be the set of predicted 3D segmentation points. To estimate the central anatomical axis, we perform a weighted Principal Component Analysis (wPCA) on $\mathcal{X}$, using the segmentation probabilities as weights. This yields a principal direction vector, along which we sample a set of candidate midline points C.

Each candidate point $c_k \in C$ is refined via a differentiable soft-snapping process. This process serves as a smooth projection mechanism, producing a refined point c_k^* using a softmax-weighted combination of all segmentation points:

$$c_k^* = \sum_{i=1}^{N} \alpha_i \, x_i, \quad \text{where} \quad \alpha_i = \frac{\exp(-\lambda \, \|x_i - c_k\|)}{\sum_{j=1}^{N} \exp(-\lambda \, \|x_j - c_k\|)}.$$

The set of refined points $\mathcal{S}_p$ defines the predicted midline.

To measure alignment between the predicted segmentation and the midline, each point $x_i \in \mathcal{X}$ is softly projected onto $\mathcal{S}_p$, producing a set of projected points $\hat{\mathcal{X}}$. These projections are subsequently used in the loss calculation to penalise deviations from the midline in a differentiable manner.

The midline loss comprises two components. The deviation loss ($\mathcal{L}_{\text{dev}}$) serves as a thickness controller by penalising the distance between each segmentation point and its nearest soft projection onto the midline:

$$\mathcal{L}_{\text{dev}} = \frac{\sum_{i=1}^{N} w_i \left(e^{\alpha \|x_i - \hat{x}_i\|} - 1 \right)}{\sum_{i=1}^{N} w_i + \epsilon}, \tag{3}$$

where w_i denotes the segmentation probabilities, and ϵ is a small constant to ensure numerical stability. A visualisation of the penalty computation pipeline for the midline loss is provided in Figure 2.

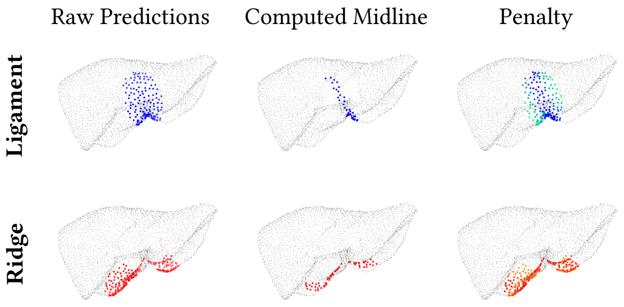

Fig. 2. Visualisation of the penalty computation pipeline for ligament and ridge regions. From left to right: raw segmentation predictions, computed midlines, and resulting deviation-based penalty maps. For ligament, penalty values range from low (blue) to high (green); for ridge, low penalties are shown in red, increasing to yellow for higher deviations from the midline.

In parallel, the alignment loss ($\mathcal{L}_{\text{align}}$) ensures that the predicted midline S_p conforms closely to the ground truth midline S_{GT}. This is accomplished by computing a soft-assigned Chamfer distance between the two midlines:

$$\mathcal{L}_{\text{align}} = \frac{1}{2} \left(\frac{1}{|S_p|} \sum_{s_p \in S_p} d_{\text{CD}}(s_p, S_{\text{GT}}) + \frac{1}{|S_{\text{GT}}|} \sum_{s_{\text{GT}} \in S_{\text{GT}}} d_{\text{CD}}(s_{\text{GT}}, S_p) \right), \tag{4}$$

where $d_{\text{CD}}(\cdot, \cdot)$ denotes the point-to-midline distance metric.

The overall midline loss is then defined as:

$$\mathcal{L}_{\text{midline}} = (1 - \lambda_m) \, \mathcal{L}_{\text{dev}} + \lambda_m \, \mathcal{L}_{\text{align}}, \tag{5}$$

with the hyperparameter λ balancing the contributions of the deviation and alignment components.

To balance the geometry-based constraints imposed by the midline loss with the region-based accuracy ensured by the weighted cross-entropy loss ($\mathcal{L}_{wCE}$), we introduce a dynamic scaling factor. This factor adjusts for the potentially different magnitudes of the losses during training:

$$\beta = \frac{\mathcal{L}_{wCE}}{\frac{1}{2} \left(\mathcal{L}_{\text{midline}}^{\text{ridge}} + \mathcal{L}_{\text{midline}}^{\text{lig}} \right) + \epsilon}, \tag{6}$$

where $\mathcal{L}_{\text{midline}}^{\text{ridge}}$ and $\mathcal{L}_{\text{midline}}^{\text{lig}}$ denote the midline losses computed for the ridge and ligament structures, respectively.

Finally, the combined total loss for the 3D segmentation task is expressed as:

$$\mathcal{L}_{\text{total}} \; = \; \alpha \, \mathcal{L}_{wCE} \; + \; \beta \left(\lambda_{\text{ridge}} \, \mathcal{L}_{\text{midline}}^{\text{ridge}} \; + \; \lambda_{\text{lig}} \, \mathcal{L}_{\text{midline}}^{\text{lig}} \right). \tag{7}$$

3.3 3D-2D Registration

3.3.1 Proposed Method. The 3D-2D registration pipeline follows an iterative optimisation approach based on differentiable rendering, similar to that proposed by the NCT team from the P2ILF challenge [Ali et al. 2025]. At each iteration, the liver mesh and its associated anatomical landmarks are transformed using the current rotation and translation parameters. The projected 3D landmarks are then directly compared against their corresponding segmented 2D landmarks to guide the optimisation process [Ali et al. 2025].

To strengthen the registration performance, the 3D mesh is augmented with an estimated silhouette contour. This silhouette is generated by tracing the shortest path along the mesh boundary, connecting anatomical extremes such as the leftmost, uppermost and rightmost points. The resulting curve is then smoothed and resampled to produce a uniformly spaced contour. This additional landmark helps provide extra guidance along the edges of the liver, where estimating depth is often difficult.

During optimisation, only the rigid transformation parameters — specifically the rotation matrix and translation vector — are updated, while the camera pose and liver scale remain fixed. The camera is placed such that it is facing the liver's anterior face, followed by a small random perturbation to simulate varied viewpoints. To improve convergence, an initial translation adjustment is performed to align the projected 3D ligament landmarks with their 2D segmentations before full optimisation begins.

The optimisation is carried out over 150 iterations using the AdamW optimiser [Loshchilov and Hutter 2019] with an initial learning rate of 0.02, alongside a plateau scheduler that adaptively reduces the learning rate based on stagnation of the loss. 30 random initialisations are performed, and the the final transformation corresponding to the lowest achieved loss is selected to ensure robustness against suboptimal starting configurations.

3.3.2 Loss Function. The loss function used to optimise registration is a weighted sum of Chamfer distances computed between the projected 3D landmarks and the segmented 2D landmarks. Chamfer distance is computed separately for the ridge, ligament, and silhouette landmarks, capturing the closest-point correspondences in the 2D image plane. Following a similar strategy to the NCT team in the P2ILF challenge, different coefficients of weights are applied to each landmark type to guide the optimisation [Ali et al. 2025]. To determine the optimal set of weights, an ablation study was conducted by varying the loss coefficients and evaluating the resulting registration accuracy using the average Chamfer distance. This process allowed us to identify the most suitable coefficients for the ligament, ridge, and silhouette landmarks.

3.4 Augmented Reality

Augmented Reality development was performed on the Varjo XR-4 Focal Edition headset. It is equipped with 90*Hz* displays providing 3840×3744 resolution at 51 pixels per degree, and 20*MP* passthrough cameras for video see-through (VST) [Varjo [n. d.]]. AR development could be performed on both Unity and through Varjo's C++ SDK, the latter was chosen as greater flexibility is required to extract the camera feed, and load deep learning models.

Firstly, an API was created from the ground up to accomplish two tasks: passthrough camera feed extraction, and object rendering onto the headset. Once completed, the camera feed can be passed to libraries such as OpenCV for processing, or the PyTorch C++ API to perform inference using our segmentation models. The rendering is performed on an OpenGL backend allowing for fine-grained control over visualisation. ImGui is used as a GUI library to modify parameters during execution.

A demonstration of the data visualisation capabilities of Augmented Reality technologies in this field was prepared, involving showing the 3D liver mesh rendering over the top of VST, with real-time conversion of the mesh to a point cloud, and rendering of landmark annotations, with the orientation of the liver able to be changed in real-time with the usage of the Varjo controllers.

4 EVALUATION

4.1 2D Segmentation

4.1.1 Datasets. Two datasets were used for the 2D segmentation task: the L3D dataset was used to initially train the model [Pei et al. 2024], and the P2ILF dataset was used for fine-tuning [Ali et al. 2025]. Three landmarks of the liver are present in the annotated masks for both datasets: the silhouette, the ridge, and the falciform ligament. The L3D dataset contains 1,152 image frames from 39 patients (122 frames are in the validation set and a further 109 make up the test set) [Pei et al. 2024]. The P2ILF dataset contains 197 frames from 11 patients (47 images from 2 patients make up the validation set and 30 images from 2 patients make up a test set) [Ali et al. 2025].

4.1.2 Evaluation Metrics. To evaluate model performance, the metrics in the P2ILF challenge 2D segmentation task are used [Ali et al. 2025], these being:

- Precision: this metric focuses on on penalising false positives to ensure correct predictions are made.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

- Dice similarity coefficient: this metric is used to ascertain the similarity between the predicted segmentation mask (Y_{pred}) and the ground truth (Y_{truth}), it is the number of predicted true positive pixels multiplied by two and divided by the sum of predicted true positive pixels and actual true positive pixels.

$$D = \frac{2 \cdot |Y_{\text{pred}} \cap Y_{\text{truth}}|}{|Y_{\text{pred}}| + |Y_{\text{truth}}|} \tag{9}$$

- Symmetric distance: Ali et al. use a symmetric distance metric proposed by François et al. [Ali et al. 2025; François et al. 2020]. $d_{\max}$ is a threshold value for whether a predicted landmark is spurious, B_I is the set of predicted image landmarks, whilst C_I is the set of ground truth image landmarks. Q is the tolerance region around the ground truth landmarks (defined by the threshold $d_{\max}$), and d_S is a symmetric distance function:

$$G = \left[\sum_{b_I \in B_I \cap Q} d_S(b_I, C_I \setminus \mathrm{FN}) + \sum_{c_I \in C_I \setminus \mathrm{FN}} d_S(c_I, B_I \cap Q) \right] \times \frac{1}{2 \cdot |C_I| \cdot d_{\max}} + \frac{|\mathrm{FP}|}{|I| - 2 \cdot |C_I| \cdot d_{\max}} + \frac{|\mathrm{FN}|}{|C_I|} \tag{10}$$

4.1.3 Experimental Setup. In training, the L3D dataset was augmented with flip transformations across both the $x-$ and $y-$ axes, followed by resizing the image to 416 × 320 pixels. The AdamW optimiser was used with a learning rate plateau scheduler [Loshchilov and Hutter 2019], which lowered the learning rate to one fifth of its original value after three epochs of no decrease in validation loss.

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

Table 1. Candidate models selected from the training ablation study. Here, Arch. is model architecture, LR is learning rate, Batch is batch size, and $\alpha/\beta/\gamma/\delta$ are the loss function hyperparameters.

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

Table 2. Fine-tuning results using the loss function $\mathcal{L}$ with $\alpha = 1.00$, $\beta = 0.00$, $\gamma = 0.00$, and $\delta = 0.00$ (see Equation 2). Highlighted cells represent a result that outperforms all P2ILF challenge teams in that metric. The best results are highlighted in bold.

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

Table 3. Fine-tuning results using the loss function $\mathcal{L}$ with $\alpha = 0.75$, $\beta = 0.00$, $\gamma = 0.00$, and $\delta = 0.25$ (see Equation 2). Highlighted cells represent a result that outperforms all P2ILF challenge teams in that metric. The best results are highlighted in bold.

A patience of 7 was used. Training was performed on an NVIDIA RTX 4070 except where VRAM was a limiting factor, in which the University of Leeds Aire HPC cluster was used, which contains GPU nodes equipped with three NVIDIA L40S GPUs per GPU node.

For the ablation study, learning rates of {0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001} are tested, and batch sizes of {2, 4, 8, 16, 32, 64} are tested (UNet3+ configurations of batch sizes greater than 8 were not feasible due to VRAM limitations). For experimentation of the loss function, 35 ablations were conducted per model on the composite loss function in Equation 2, the hyperparameters (α, β, γ, and δ) for the loss function selected from these experiments can be seen in the fifth column of Table 1.

4.1.4 Quantitative Results. Table 1 shows 10 candidate models selected for fine-tuning from the ablation study. At least one model was picked from each architecture, alongside other models that performed well in the evaluation. To ensure fairness, all P2ILF 2D segmentation results were evaluated on our metrics before comparison (Available in appendix in Table C1). Table 3 shows the evaluation results of the 10 chosen candidate models, prior to fine-tuning, after fine-tuning, and after post-processing was applied. This resulted in half of the models beating or matching the best results in P2ILF in every metric, with 3 models outperforming in every metric outright.

The best performing candidate model was Candidate 2, providing an 11% increase (30% relative increase) in mean precision, a 5% increase (over 13% relative increase) in mean Dice score, and a 6% decrease (over 13% relative improvement) in mean symmetric distance, compared to the teams that competed in the P2ILF challenge (see Table 3).

During inference, the time taken to infer and post-process masks was recorded. An inference ranged from 9 to 15 milliseconds, whilst post-processing added a time penalty ranging from 9 to 12 milliseconds.

4.1.5 Qualitative Results. Figure 3 shows six examples of images from the P2ILF test set (3 images from patient 4 and 3 images from patient 11) [Ali et al. 2025], annotated with masks. Sub-figure A is the ground truth, Sub-figures B and C are from the best performing

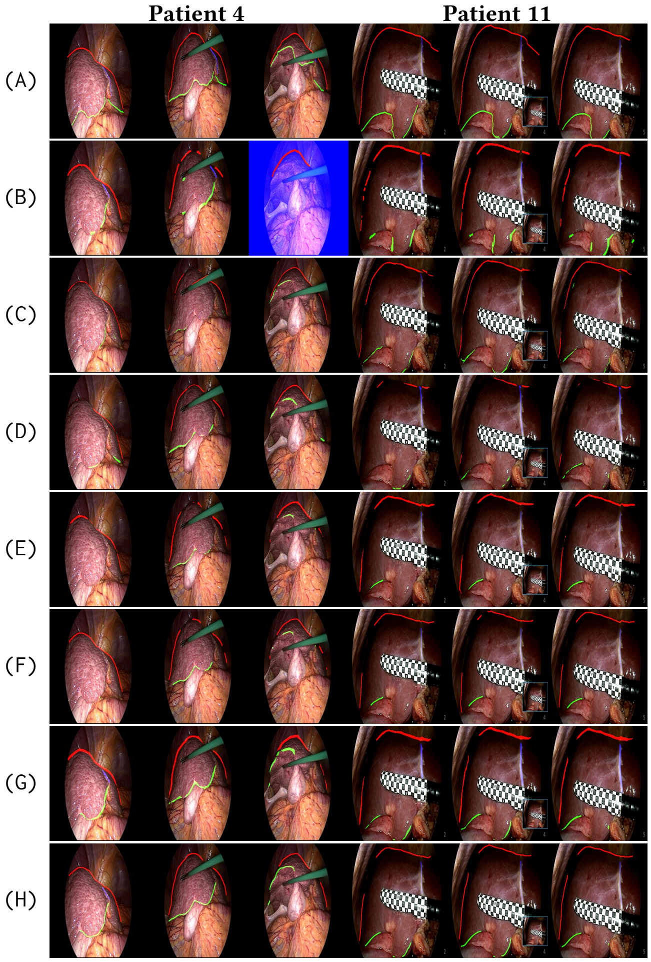

Fig. 3. Qualitative results on the P2ILF test set [Ali et al. 2025], with (A): ground truth, (B): BHL predictions, (C) NCT predictions, (D) Candidate 2 initial training (E) Candidate 2 CE fine-tune, (F) Candidate 2 CE fine-tune with post-process, (G) Candidate 2 CE+FTL fine-tune, (H) Candidate 2 CE+FTL fine-tune with post-process.

teams in the P2ILF challenge 2D segmentation task (BHL and NCT respectively) [Ali et al. 2025]. It can be seen that both models under-predict, and false positive pixel classification occurring when tools are present in the image. Sub-figures D through H all show variations of Candidate 2, the most performant candidate model. Sub-figure D shows Candidate 2 after initial training, prior to any fine-tuning. This model also under-predicts (low recall), but qualitatively, rarely makes false positive predictions (high precision); noticeably, the ligament is not predicted on the images used from patient 4. Sub-figure E shows Candidate 2 after being fine-tuned with the P2ILF training set with solely cross-entropy loss, contours remain thin and the model still under-predicts, landmark contours (especially the ridge) are disconnected, the ligament on patient 4 remains unpredicted. Sub-figure F shows the model used in sub-figure E after post processing, Sub-figure G shows Candidate 2's predictions after fine-tuning with a combination of cross-entropy loss and Focal-Tversky loss, but prior to mask post-processing. This model predicts landmarks more often than the aforementioned models, but at the cost of thickening contours; this promotion of making predictions now means that the ligament of patient 4 is now predicted in the left-hand image. Finally, sub-figure H shows the cross-entropy and Focal-Tversky loss fine-tuned Candidate 2 model, with our novel mask post-processing applied. Contours are now of a consistent thickness, having thinned out the thicker contours present in sub-figure E. Unlike the other models shown, sub-figure H does not show contours with frequent disconnections, forming clean contours around the liver landmarks.

4.2 3D Segmentation

4.2.1 Datasets. To ensure the robustness and generalisation of our proposed method, we used two publicly available datasets of 3D liver models, including a large dataset obtained from Zhang et al. that combines three public datasets (3Dircadb [Soler et al. 2010], LiTS [Bilic et al. 2023], and Amos [Ji et al. 2022]) [Zhang et al. [n. d.]]. The other smaller dataset, consisting of 9 training and 2 test patients, was sourced from the P2ILF challenge [Ali et al. 2025]. Data for each patient includes a 3D liver model, saved as a wavefront object, and an XML file containing the anatomical annotations. These files are parsed to extract detailed contour information of the type of anatomical structure (ridge or ligament) and the corresponding indices of mesh vertices that define these contours. These extracted annotations were then used to create numerical labels for the 3D mesh vertices.

4.2.2 Evaluation Metrics. We assess the performance of our segmentation framework using 3D Chamfer distance [Huang et al. 2023], which quantifies the point-to-point average distance between the segmented landmarks and the ground truth. It functions by averaging the minimum distance between points in two point clouds. For two point sets X and Y, it is defined as:

$$d_{\text{CD}}(X, Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}\|x-y\|_p + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}\|y-x\|_p, \quad (11)$$

where $\|\cdot\|_p$ denotes the p-norm distance of the points.

4.2.3 Experimental Setup. Liver meshes were converted into point cloud arrays for compatibility with PointNet++, typically containing between 4000 and 15000 points. These models were standardised by either furthest point sampling or padding the vertices to achieve a fixed number of 4096 points. The labelled point clouds were then encapsulated in compressed files that store the vertex coordinates and their labels.

The dataset is partitioned into training, validation and test sets in a 60:20:20 split, with data normalised for consistency purposes. Class weights are computed using the inverse square root class frequency weighting method, assigning higher weights to classes with lower frequencies. This is necessary due to heavy class imbalance, with low frequency in the ligament class compared to the liver class. We augment the dataset to increase the diversity of the training data to promote generalisation, improving its performance on unseen samples. Three different augmentations were performed, including upscaling, downscaling and rotations on the z-axis. Point clouds are randomly scaled in the range of 65%–145% of their original size, and then randomly rotated in the range of $-180°$-$180°$on the z-axis.

To systematically evaluate the impact of the learning rate and loss function components (Equation 7) on segmentation performance, a structured grid search was performed. Each configuration is trained independently, and the resulting segmentation quality is evaluated using 3D Chamfer distance [Huang et al. 2023]. Learning rate selection is performed separately through another dedicated ablation experiment. The evaluated learning rates ranged from 0.00025 to 0.01, examining their effects on convergence stability and final segmentation accuracy. All experiments were conducted on an NVIDIA RTX 4070. A batch size of 32 was set and an AdamW optimiser with default parameters was utilised for training [Loshchilov and Hutter 2019]. A learning rate scheduler was implemented to reduce the rate once the validation loss plateaued and early stopping was applied if the validation loss remained stagnant.

In our experiments, negative log-likelihood (NLL) loss was adopted as the baseline due to its use as the standard loss in PointNet [Qi et al. 2017a] and PointNet++ [Qi et al. 2017b]. We also evaluated the weighted cross-entropy (wCE) loss, which incorporates the softmax operation internally. Although both losses yield similar outcomes when properly configured, the slight differences in their implementation can influence convergence behaviour and final accuracy, making the inclusion of wCE a valuable alternative for comparison.

The optimal hyperparameters were selected based on performance on the validation dataset. For the combined dataset [Zhang et al. [n. d.]], the best-performing configuration used a weighted cross-entropy weight of $\alpha = 0.25$, with geometry-based ridge and ligament losses set at $\lambda_{\text{ridge}} = 0.5$ and $\lambda_{\text{lig}} = 0.25$, respectively. A learning rate of 0.0075 was chosen, as it yields the lowest Chamfer distances. For the P2ILF dataset, the best configuration selected used $\alpha = 0.25$, $\lambda_{\text{ridge}} = 0.75$, and $\lambda_{\text{lig}} = 0$, with an optimal learning rate of 0.0005.

4.2.4 Quantitative Results. The quantitative results recorded during the series of ablations are presented in Table 4 and Table 5. Table 4 presents our findings when our models are tested on the P2ILF challenge dataset [Ali et al. 2025], and Table 5 shows our findings tested on Zhang et al.'s combined dataset [Zhang et al. [n. d.]]. Furthermore, we include the results of the top two performing teams from the P2ILF challenge paper as a means of comparison against other literature. The two teams included from the P2ILF challenge are: UCL, which achieved the best results on the ligament, and NCT, which achieved the best results on the ridge. Each model is tested on the hold-out set of the dataset it was trained on, as well as the other dataset to verify generalisation capabilities.

Table 5 indicates that when testing on the combined dataset, the two best performing configurations of PointNet++ with weighted cross-entropy loss and PointNet with negative loss likelihood loss achieve the lowest mean Chamfer distance and ultimately outperform the best performing configurations from the P2ILF challenge [Ali et al. 2025]. However, it should be noted that despite the high performance on the combined dataset, these configurations generalise poorly to the P2ILF dataset, which can be observed in Table 4. For instance, the PointNet++ configuration using weighted cross-entropy loss records a mean Chamfer distance of 12.71*mm* on the combined dataset but deteriorates sharply to 41.99*mm* when generalising to the P2ILF dataset. Similarly, the PointNet model with NLL loss (LR = 0.00075) achieves a competitive 12.27*mm* on the combined dataset, yet it fails to generalise, reaching 199.94*mm* on the P2ILF dataset.

	Model	Loss	LR	Train	ch_r	ch_l	Mean
	PointNet++	Midline, wCE	0.01	‡*	**19.70**	**13.46**	**16.58**
	PointNet++	NLL	0.0005	P2ILF	22.23	55.77	39.00
	PointNet++	wCE	0.01	‡*	23.72	60.26	41.99
	PointNet++	wCE	0.0005	P2ILF	38.43	69.62	54.03
	PointNet++	Midline, wCE	0.00075	P2ILF	36.11	116.10	76.11
	PointNet	NLL	0.01	P2ILF	95.85	73.46	84.65
	PointNet++	NLL	0.005	‡*	F	115.54	115.54
	PointNet	NLL	0.00075	‡*	199.94	F	199.94
Teams [Ali et al. 2025]							
UCL	PointNet++	NLL, HFD	0.001	P2ILF	27.97	24.47	26.22
NCT	2× MeshCNN	wCE	NA	P2ILF	27.19	36.38	31.79

Table 4. **Evaluation on the P2ILF challenge test dataset [Ali et al. 2025].** LR: learning rate. ch_r, ch_l: Chamfer distances (in *mm*) for ridge and ligament, respectively. ‡: trained on combined dataset [Zhang et al. [n. d.]]. * indicates evaluation on the unseen test set. Highlighted cells represent a metric result that outperforms all P2ILF challenge teams. The best results are in bold.

Model	Loss	LR	Train	ch_r	ch_l	Mean
PointNet	NLL	0.00075	‡	7.92	**16.62**	**12.27**
PointNet++	wCE	0.01	‡	**7.13**	18.30	12.71
PointNet++	Midline, wCE	0.01	‡	11.62	21.57	16.60
PointNet++	Midline, wCE	0.00075	P2ILF*	19.50	23.21	21.35
PointNet++	NLL	0.0005	P2ILF*	28.79	17.95	23.37
PointNet	NLL	0.01	P2ILF*	27.39	29.10	28.25
PointNet++	wCE	0.0005	P2ILF*	33.49	23.48	28.49
PointNet++	NLL	0.005	‡	32.21	30.39	31.30

Table 5. **Evaluation on the combined test dataset [Zhang et al. [n. d.]].** LR: learning rate. ch_r, ch_l: Chamfer distances (in *mm*) for ridge and ligament. ‡: trained on combined dataset. *: evaluated on unseen test set. The best results are in bold.

Furthermore, Table 4 shows that PointNet++ with combined midline and weighted cross-entropy loss functions along with a learning rate of 0.01, when trained on the combined dataset and tested on the P2ILF dataset, outperforms all teams from the P2ILF challenge and achieves the best results out of all our models tested on the P2ILF data. Although the Chamfer distances reported by this model are lower than those presented in Table 5, they nonetheless demonstrate that our proposed method generalises effectively to unseen data, as evidenced by its performance on the hold-out set. For example, this method, using the midline loss at LR = 0.01, reduces the mean Chamfer distance to 16.58*mm* — a reduction exceeding 36% compared to the UCL team's method.

4.2.5 Qualitative Results. Figure 4 provides a comparative visual analysis of segmentation outputs from models trained on the combined dataset for four patients. Two from the P2ILF test set (patient

4 and patient 11) [Ali et al. 2022] and two from the combined dataset (LiTS-65 and Amos-119) [Zhang et al. [n. d.]]. Figure 5 is also included as a comparison between the top performing teams in the P2ILF challenge and our best model.

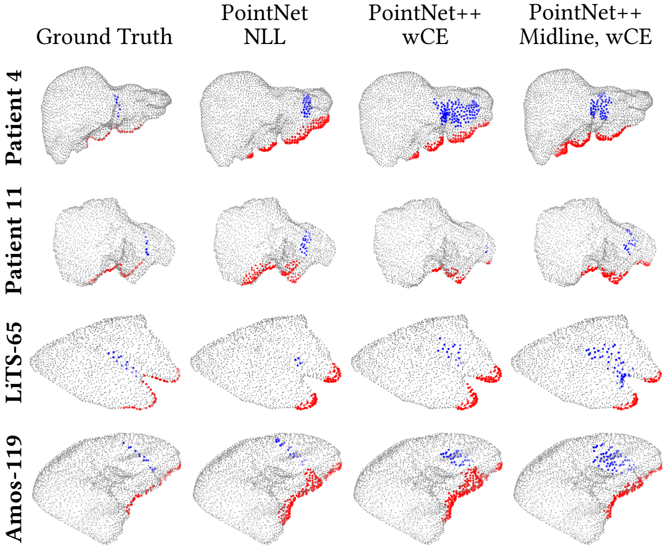

Fig. 4. Qualitative comparison of segmentation results between our proposed method and baselines from P2ILF [Ali et al. 2022] (patients 4 and 11) and the combined dataset [Zhang et al. [n. d.]] (LiTS-65 [Bilic et al. 2023] and Amos-119 [Ji et al. 2022]). Ligament points are coloured blue; Ridge points are coloured red.

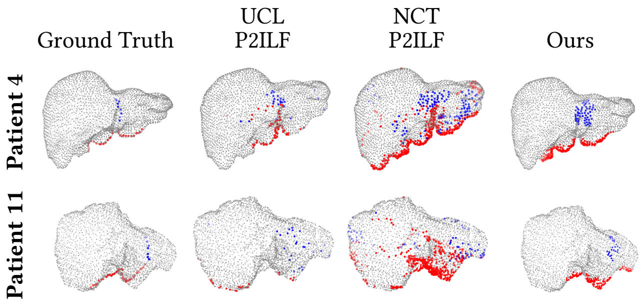

Fig. 5. Qualitative comparison of segmentation results between the two best teams from the P2ILF challenge [Ali et al. 2022] and our best model. Test patients 4 and 11 are used. Ligament points are coloured blue; Ridge points are coloured red. Points on the backside of the liver are displayed with a lower alpha value.

The results indicate that the segmentation produced with the midline loss is more closely bound to the ground truth. Notably, in both patient 4 and patient 11 in Figure 4, the ridge structure is more accurately delineated as the predicted midline closely curves around the ridge, in contrast to the broader, region-like segmentations observed with the weighted cross-entropy model. Although the overall segmentation size remains relatively large for the ligament, the localisation is notably improved with the midline loss, showing a more generalised and anatomically consistent alignment.

These improved results are also evident in the comparisons made in Figure 5. Segmentations produced by our model demonstrate improvements over the two highest scoring P2ILF teams (Table 4). In particular, the predicted ridge and ligament regions are more accurately aligned with the ground truth in contrast to the P2ILF predictions, which are erratic and incorrectly placed some points behind the liver.

By incorporating a geometric aspect, the midline loss improves localisation by effectively carving the ridge around the liver while ensuring thin segmentation. Similar improvements in localisation and accuracy are also observed on the dataset provided by Zhang et al., particularly for the ridge segmentation.

4.3 3D-2D Registration

4.3.1 Evaluation Metrics. The evaluation metric used in the 3D-2D registration task is reprojection error, this is the measure of the distance between a set of projected points (i.e. our projected landmarks) and the ground truth. Following in the steps of Ali et al., we utilise the Hausdorff distance metric to measure reprojection error for the ridge and ligament landmarks [Ali et al. 2025]. The reprojection error, R, can be defined as:

$$R = \frac{1}{|C_M|} \sum_{c_M} d_H(\Pi(c_M), c_I), \tag{12}$$

where C_M is the set of ground-truth landmark points, C_I is the set of vertex projections, d_H represents the Hausdorff distance, and Π is our 3D-2D registration (projection) function.

4.3.2 Quantitative Results. The landmark coefficients used in our method for the ridge, ligament, and silhouette landmarks were 0.5, 1, and 0.5, respectively. This was chosen from the ablation study as it produced the lowest total loss during optimisation (See Appendix Table C2 for further detail). The ligament was emphasised due to its strong influence on the liver's orientation while covering the least area [Ali et al. 2025].

Table 6 shows the reprojection errors in pixels for ridge (`rpe_r`) and the falciform ligament (`rpe_l`) landmarks across patient 4 and 11. Our proposed method outperforms the best performing team in the P2ILF challenge (NCT) [Ali et al. 2025] across the majority of cases. Notably, a substantial reduction in ligament error (`rpe_l`) is observed for patient 4, where our model yields errors in the range of 17.03–110.89 pixels, whereas previous methods report significantly higher errors exceeding 400 pixels. This highlights the effectiveness of our ligament-guided optimisation in scenarios where the anatomy is relatively well preserved. In contrast, reprojection errors for patient 11 are comparatively higher to patient 4. The ridge and ligament landmarks in these images show increased variation, primarily due to the visual complexity of the liver during laparoscopic procedures. Such cases pose a greater challenge for rigid registration methods.

Image	P2ILF Avg. [Ali et al. 2025]		NCT [Ali et al. 2025]		Ours	
	rpe_r	rpe_	rpe_r	rpe_l	rpe_r	rpe_l
4_3	565.27	566.04	401.36	257.95	**242.32**	**74.03**
4_4	705.67	643.17	494.53	368.75	**194.65**	**77.49**
4_7	509.32	395.07	**115.73**	170.76	247.36	**110.89**
4_11	596.26	612.36	360.19	329.40	**127.44**	**74.06**
4_17	524.41	1076.29	323.60	458.22	**110.48**	**17.03**
4_20	608.47	611.06	**183.58**	393.21	211.72	**21.95**
4_21	567.00	—	159.30	—	**156.16**	—
4_22	550.47	—	212.36	—	**110.61**	—
11_2	985.24	899.17	1008.61	**356.36**	**614.47**	**319.79**
11_3	804.91	1085.95	842.67	**177.02**	**598.03**	292.25
11_4	854.54	531.91	720.35	**185.74**	**447.03**	326.91
11_5	813.15	572.49	788.44	**311.52**	**484.11**	326.66
11_6	928.90	582.23	807.11	543.89	**476.27**	**340.11**
11_7	889.76	798.65	360.03	**408.01**	**443.85**	**297.38**
11_8	790.64	536.70	**329.80**	**237.32**	419.09	336.92
11_9	707.74	541.87	**247.95**	**270.65**	419.26	335.89
Mean	710.92	675.21	466.80	319.20	**331.43**	**210.81**
Overall	693.07		393.00		**271.12**	

Table 6. Reprojection errors in pixels from the two test patients. These errors are calculated for the ridge (rpe_r) and the falciform ligament (rpe_l) by comparing the projected 3D ground-truth landmark vertices from the registered model with the corresponding 2D ground-truth pixel locations. NA indicates cases with missing data. The overall mean is obtained by averaging rpe_r and rpe_l values.

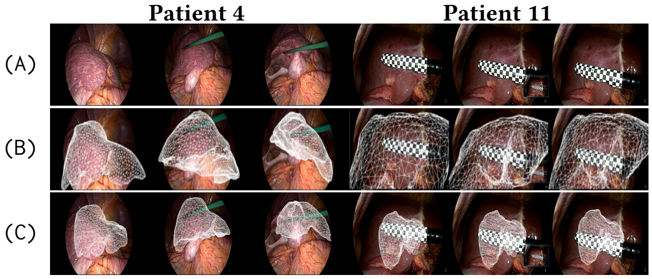

Fig. 6. Qualitative results for patients 4 and 11.(A) Original laparoscopic images, (B) NCT team's registration results from P2ILF [Ali et al. 2025] and (C) our registration model results

4.3.3 Qualitative Results. Figure 6 presents qualitative comparisons for patient 4 and 11. For patient 4, the visual overlay of the registered 3D liver mesh onto the 2D image demonstrates accurate spatial alignment. Compared to the NCT's method from the P2ILF challenge [Ali et al. 2025], our model more closely adheres to the anatomical boundaries, particularly around the falciform ligament and ridge structures. In several cases, the NCT's overlay shows clear misalignment or excessive deformation, often extending beyond the actual liver boundaries in the laparoscopic image, indicating poor spatial alignment. In contrast, our model produces a more anatomically coherent fit that respects image boundaries whilst accurately tracing liver contours. For patient 11, our method struggles with precise alignment due to the limited visibility of the liver in the image. Despite this, our model maintains a reasonable fit by aligning the general liver shape and orientation.

5 DISCUSSION

5.1 2D Segmentation

Table 2 and Figure 3 sub-figure E show that the fine-tuned model, with only cross-entropy loss, is precise when predicting, however it is visible that the recall of the model is low. Due to this under-prediction, the benefits of post-processing are minimised (see sub-figure F) due to the thin and sparse landmark predictions outputted by the model compared with sub-figures G and H. Table 2 verifies this, showing much smaller improvements between `tune` and `post` (and lower overall values) when compared to Table 3.

Table 3 and sub-figure G from Figure 3 show a model that is less prone to under-prediction when compared to sub-figure E, predicting a greater proportion of the landmarks, albeit with thicker contours. This is due to the Focal-Tversky loss component promoting landmark predictions, with the focal component adding emphasis to predictions that are of low certainty (i.e. difficult predictions). Post-processing (sub-figure H) mitigates over-prediction by thinning the contours, as well as connecting nearby components, the contours are also smoothed. Post-processing improves the quality of the contours, as shown by the increase in the DSC and symmetric distance scores in Table 3.

In regards to future direction, Pei et al.'s approach of estimating depth masks focuses on providing clearer information to the CNN, such that it can predict with greater accuracy [Pei et al. 2024].

Labrunie et al. utilise distance maps as opposed to binary masks when training their 3D-2D registration model [Labrunie et al. 2024], an approach that can be applied to 2D segmentation. The mask post-processing presented in this report lays out the fundamentals of this approach, with many other aspects yet to be explored, such as mask correction models. There is a question of whether DSC is an appropriate metric, due to the linear nature of the landmarks; alternatives such as centre line Dice (clDice) should have their suitability investigated in this field [Shit et al. 2021].

5.2 3D Segmentation

Our experimental results provide compelling quantitative evidence for the efficacy of our proposed method in 3D liver landmark segmentation. By leveraging a significantly larger combined dataset of 300 unique 3D liver models and incorporating our geometry-constrained midline loss, our PointNet++ model achieves mean Chamfer distances of 16.60*mm* on the combined dataset and 16.58*mm* on the P2ILF dataset. These outcomes highlight the limitations of earlier methods, particularly those from the P2ILF challenge teams, which were constrained by limited training data and struggled to generalise. The best results on the combined dataset, even when using a model trained solely on the smaller P2ILF data, demonstrate the improvements in generalisations achieved by our method. The incorporation of the midline loss is particularly important for enabling robust performance across varied liver datasets. As shown in Table 4, the PointNet++ model with midline and weighted cross-entropy loss, when trained on the smaller P2ILF dataset with a learning rate of 0.00075, still performed well on the combined dataset, achieving a mean Chamfer distance of 21.35*mm* compared to 16.60*mm* for the model trained on the larger dataset with a learning rate of 0.01

Furthermore, visual comparisons reveal that segmentation outputs using the midline loss exhibit a distinct, precise curvature of the ridge and are more tightly aligned with the ground truth. In contrast, predictions from models trained with wCE loss tend to be broader and less defined, resembling region-like segmentations. The improved localisation achieved by carving the ridge around the liver confirms the enhanced geometric fidelity of our method and demonstrates its ability to generalise to unseen data from different datasets. This is especially evident when comparing performance on the combined dataset to the P2ILF dataset, where our method substantially reduces the mean Chamfer distance and outperforms previous configurations. Overall, the combination of a robust training dataset and a geometry-aware loss function yields improvements in both quantitative metrics and qualitative assessments, providing a strong foundation for future advancements in 3D liver segmentation.

In terms of future research, the paper on deep point-graph representation [Xie et al. 2025] leverages a hybrid point-graph representation that combines point-based learning with graph neural networks and an implicit surface decoder to preserve intricate anatomical topology. This method ensures that connectivity and fine surface details are maintained, an aspect that could greatly enhance the geometric fidelity of liver segmentations. In contrast, WS-TIS [Wang et al. 2024] adopts a weakly supervised, two-stage approach that couples multi-label classification with instance segmentation. By efficiently utilising limited annotations through feature disentanglement and gated attention mechanisms, it achieves robust segmentation accuracy with reduced labelling requirements. Both techniques, although applied to different anatomical contexts, offer promising strategies that could be implemented in the liver segmentation framework to improve both precision and generalisability in capturing complex liver structures.

5.3 3D-2D Registration

Table 6 and Figure 6 compare our registration approach against the top-performing team, NCT, in the P2ILF challenge [Ali et al. 2025]. For patient 4, the proposed method achieves substantially lower reprojection errors, particularly for the ligament landmarks, where we observe a decrease in error from over 400 pixels (as reported by NCT) to a range between 21.95–110.89 pixels. This improvement is a direct result of the ligament-guided optimisation, which helps preserve anatomical structure during alignment, especially when the liver appears relatively undeformed in the laparoscopic image. Figure 6 sub-figure C shows that our projected mesh fits more tightly within the visible liver region compared to the NCT prediction (sub-figure B), which often overshoots the anatomical boundaries. For patient 11 however, the registration proves more difficult; Table 6 shows higher errors across both ridge and ligament landmarks. This is due to the poor visibility observed in patient 11's laparoscopic scenes, which makes registration difficult.

To address this, future work could incorporate deformable registration frameworks, such as FEA [Labrunie et al. 2024], to simulate non-rigid liver behaviour under surgical conditions. This would enable more realistic transformations, especially in cases of extreme liver deformation. Notably, our method includes silhouette-based supervision during optimisation, which improves alignment in the *z*-direction. Unlike ridge and ligament, silhouette landmarks provide extensive spatial information, offering greater insight into the overall shape, which helps resolve the depth ambiguity common in 3D-2D registration. As shown in sub-figure C, this results in a better spatial fit, particularly around the liver's overall shape.

6 CONCLUSION

We presented a fully automated pipeline for augmenting laparoscopic liver surgery with machine-assisted segmentation and registration. Our 2D segmentation framework evaluated four UNet variants and DeepLabV3+ with a composite loss. A novel post processing step removed low confidence false positives, applied smoothing and used skeletonisation to refine anatomical landmarks. Trained on the L3D dataset and fine-tuned on P2ILF images, the best model achieved up to a 30% relative increase in precision and 13% improvement in Dice score and symmetric distance over the P2ILF challenge's leading teams.

In 3D segmentation, we used PointNet++ with point clouds of preoperative liver meshes and introduced the innovative midline loss. On both the combined public dataset and the P2ILF test set, our methodology reduced mean Chamfer distances by over 36% compared to previous baselines, demonstrating superior generalisation to unseen livers.

We then employ an iterative differentiable rendering registration procedure that refines rigid pose parameters by minimising Chamfer distances between the projected 3D landmarks to their 2D counterparts. We perform 30 random initialisations and select the optimal fit based on the validation loss. This approach achieves a reduction of over 31% in mean reprojection error compared to leading P2ILF methods.

Finally, we implemented the pipeline on a Varjo XR-4 headset using a custom C++ SDK to extract camera footage, and OpenGL to perform 3D rendering.

ACKNOWLEDGMENTS

We would like to thank our project supervisor, Dr. Sharib Ali, for his continued guidance and insights throughout the duration of this project, as well as supporting our submission of two papers to the 29th Medical Image Understanding and Analysis conference (MIUA 2025). We would like to extend this thanks to fellow supervisor and assessor, Dr. Rafael Kuffner dos Anjos, for granting us access to the Virtual Reality lab and for his advice. Finally, we would like to express our gratitude to Tao Chen and Xukun Zhang for sharing knowledge in many aspects of the project.

REFERENCES

Yinoussa Adagolodjo, Raffaella Trivisonne, Nazim Haouchine, et al. 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)*. 539–544.

Sharib Ali, Yamid Espinel, Yueming Jin, et al. 2025. An objective comparison of methods for augmented reality in laparoscopic liver resection by preoperative-to-intraoperative image fusion from the MICCAI2022 challenge. *Medical Image Analysis* 99 (2025), 103371. https://doi.org/10.1016/j.media.2024.103371

Sharib Ali, Yueming Jin, Yamid Espinel Lopez, and Adrien Bartoli. 2022. Preoperative to intra-operative laparoscopic fusion challenge.

Jorge Bernal, F. Javier Sánchez, and Gloria Fernández-Esparrach and. 2015. *CVC-ClinicDB*.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, et al. 2023. The Liver Tumor Segmentation Benchmark (LiTS). *Medical Image Analysis* 84 (2023), 102680.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, et al. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, 833–851.

Jia Deng, Wei Dong, Richard Socher, et al. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

David H Douglas and Thomas K Peucker. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization* 10, 2 (1973), 112–122.

Tom François, Lilian Calvet, Sabrina Madad Zadeh, et al. 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, 7 (01 Jul 2020), 1177–1186.

Rana Hanocka, Amir Hertz, Noa Fish, et al. 2019. MeshCNN: a network with an edge. *ACM Trans. Graph.* 38, 4, Article 90 (July 2019), 12 pages.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, et al. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

Huimin Huang, Lanfen Lin, Ruofeng Tong, et al. 2020. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1055–1059.

Tianxin Huang, Qingyao Liu, Xiangrui Zhao, et al. 2023. Learnable Chamfer Distance for Point Cloud Reconstruction. arXiv:2312.16582 [cs.CV]

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, et al. 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*. 225–2255.

Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, et al. 2020. Kvasir-SEG: A Segmented Polyp Dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II* (Daejeon, Korea (Republic of)). Springer-Verlag, Berlin, Heidelberg, 451–462.

Yuanfeng Ji, Haotian Bai, Chongjian GE, et al. 2022. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 36722–36732.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, et al. 2018. End-to-End Recovery of Human Shape and Pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7122–7131.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. 2023. Segment Anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 3992–4003.

Bongjin Koo, Maria R. Robu, Moustafa Allam, et al. 2022. Automatic, global registration in laparoscopic liver surgery. *International Journal of Computer Assisted Radiology and Surgery* 17, 1 (01 Jan 2022), 167–176.

Mathieu Labrunie, Daniel Pizarro, Christophe Tilmant, et al. 2024. Automatic 3D/2D Deformable Registration in Minimally Invasive Liver Resection using a Mesh Recovery Network. In *Medical Imaging with Deep Learning (Proceedings of Machine Learning Research, Vol. 227)*. PMLR, 1104–1123.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. 2009. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision* 81, 2 (01 Feb 2009), 155–166. https://doi.org/10.1007/s11263-008-0152-6

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Islem Mhiri, Daniel Pizarro, and Adrien Bartoli. 2025. Neural patient-specific 3D–2D registration in laparoscopic liver resection. *International Journal of Computer Assisted Radiology and Surgery* 20, 1 (01 Jan 2025), 57–64. https://doi.org/10.1007/s11548-024-03231-x

Jialun Pei, Ruize Cui, Yaoqian Li, et al. 2024. Depth-Driven Geometric Prompt Learning for Laparoscopic Liver Landmark Detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Springer Nature Switzerland, Cham, 154–164.

Charles R. Qi, Hao Su, Kaichun Mo, et al. 2017a. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Charles Ruizhongtai Qi, Li Yi, Hao Su, et al. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/d8bf84be3800d12f74d8b05e9b89836f-Paper.pdf

Urs Ramer. 1972. An iterative procedure for the polygonal approximation of plane curves. *Computer graphics and image processing* 1, 3 (1972), 244–256.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, Cham, 234–241.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, et al. 2021. clDice - A Novel Topology-Preserving Loss Function for Tubular Structure Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16560–16569.

Douglas P Slakey, Eric Simms, Barbara Drew, et al. 2013. Complications of liver resection: laparoscopic versus open procedures. *JSLS* 17, 1 (Jan. 2013), 46–55.

Luc Soler, Alexandre Hostettler, Vincent Agnus, et al. 2010. 3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep* 1, 1 (2010).

Varjo. [n. d.]. Mixed Reality Headset for Professionals. https://varjo.com/products/xr-4/

Haoyu Wang, Kehan Li, Jihua Zhu, et al. 2024. Weakly supervised tooth instance segmentation on 3d dental models with multi-label learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 723–733.

Kangxian Xie, Jiancheng Yang, Donglai Wei, et al. 2025. Efficient anatomical labeling of pulmonary tree structures via deep point-graph representation-based implicit fields. *Medical Image Analysis* 99 (1 2025), 103367.

Xu Yan. 2019. Pointnet/Pointnet++ Pytorch. (2019).

Zhiding Yu, Chen Feng, Ming-Yu Liu, et al. 2017. CASENet: Deep Category-Aware Semantic Edge Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 1761–1770.

T. Y. Zhang and C. Y. Suen. 1984. A fast parallel algorithm for thinning digital patterns. *Commun. ACM* 27, 3 (March 1984), 236–239.

Xukun Zhang, Sharib Ali, Minghao Han, et al. [n. d.]. Two-stream MeshCNN for key anatomical segmentation on the liver surface. *International Journal of Computer Assisted Radiology and Surgery* ([n. d.]). In press.

Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. 2018. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters* 15, 5 (2018), 749–753.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, et al. 2020. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging* 39, 6 (2020), 1856–1867.

A PROJECT MANAGEMENT

A.1 Supervisor and Theme

Once our group was formed, it was decided that the requirements for a project was not found in a specific topic, but in desirable aspects within the project and area of research. The primary requirement was that of producing an impactful piece of work, with a view to possibly publish any novel methods and findings. Many potential supervisors were contacted and discussed with, with varying levels of agreement within the group on the suitability of their field of research. Once we met with Dr. Sharib Ali, he prepared a presentation and proposed that we worked on his ARMADILLO project, originating as a challenge he hosted at MICCAI 2022. ARMADILLO is an end-to-end 3D-2D liver registration pipeline that utilises segmentation models and a 3D-2D registration solution reliant on segmentation outputs, combined with a hardware implementation (Varjo XR-4 FE headset). The scale of this project was not understated, but it sparked interest in the entire group with the project's novelty and potential impact, leading us to go ahead with his project.

The group met with Sharib once every two weeks, changing to once every week as we got closer to the MIUA submission deadline. After the initial meetings where our understanding of the technical aspects of the project was cemented, each meeting consisted of presenting developments in each of the four areas of work: 2D segmentation, 3D segmentation, 3D-2D registration, and Augmented Reality development. This would then allow any queries to be asked, whilst also making clear in which areas the project was thriving or struggling.

A.2 Group organisation

The group was organised into four sub-groups, one for each task. This consisted of teams for 2D segmentation, 3D segmentation, 3D-2D registration, and augmented reality. 2D segmentation consisted of Jibran and Abhinav; 3D segmentation consisted of Karim and Aodhan; 3D-2D registration consisted of Najmi and James; augmented reality was developed by Theodora with assistance from James. At a later date James also worked on 2D segmentation. Each team followed an agile-like development strategy wherein each team conducted background research into relevant literature and associated methods, before implementing and testing these methods for suitability in their respective tasks. Once we reviewed quantitative and qualitative evidence, new methods were chosen for exploration or successful methods were refined further with novel additions added to develop successful approaches to each task, such as 3D segmentation's midline loss or 2D segmentation's post-processing.

Group communication was primarily done in-person, as many members worked on campus almost daily due to the scale of the project. Outlook was used to schedule formal meetings, with informal communication occurring over mobile devices. A GitHub organisation was created for the project, with each task having its own repository within the organisation, and development occurring in the form of pair programming or individually within close vicinity of one another. Consequently, Git was used as a form of version control for coding aspects within the project, with members on each team working on separate branches for each feature and updating changes regularly during development. Overleaf was used in writing the group report, individual reports and the poster to provide a shared writing space and also to provide version control for writing aspects of the project.

Meetings occurred every two weeks for the majority of the project. Initial meetings revolved around the analysis and discussion of researched methods that could be used for each task. This formed the basis of our understanding, with our initial experiments and approaches to tasks being decided here. As time progressed, meetings shifted focus to the progress we made and results we could demonstrate, with queries being asked to our supervisor to further progress on each task. At these meetings, it became clear what approaches would work and which approaches would not work, as well as any necessary adjustments to be made. This style of meeting happened more frequently towards the end of the project to ensure our work was of the highest standard. At each opportunity, we utilised our supervisor's expertise and contacts within the field. This included virtual meetings with post-doc students from around the world. This was especially useful in early stages of the project, when it came to contacting authors of relevant literature for further clarifications on their works and accessing pre-published work. In the latter stages of the project, our supervisor was able to provide data from teams participating in the P2ILF challenge, from which we could compare our own quantitative and qualitative results.

These meetings ensured every member of the group was regularly updated with the latest developments both in code and in report writing. Tasks were also assigned to each member during these meetings, in the presence of our supervisor, that were often due for the next scheduled meeting. Through this system of regular discussion of updates and regular goal-setting, each member of the group was kept on track and it was ensured that deadlines were met on time. Care was also taken at each meeting to make sure that each member was assigned a task of similar workload, such that nobody was without work and such that each member had an manageable amount of work. In the rare case that one team struggled to complete a task on time, other group members helped out where possible to ensure the smooth development of the project.

A.3 Planning and execution

Figure A1 shows the original project timeline, while Figure A2 shows the timeline as it actually occurred. Although the two charts align closely in most respects, the main deviation occurred when the opportunity to submit two papers to the MIUA conference presented itself towards the end of the project. Preparing this submission demanded a significant investment of time and resources, which necessitated a temporary reallocation of effort away from some planned tasks.

Despite this disruption, we were able to adjust our schedule and recover lost time without compromising the overall deadline date. In particular, the decision to split the team into sub-groups, one focusing on 2D segmentation, another on 3D segmentation, and a third on the early stages of registration proved very valuable. By running these teams in parallel, we retained enough flexibility to deal with the additional workload. As a result, not only did we

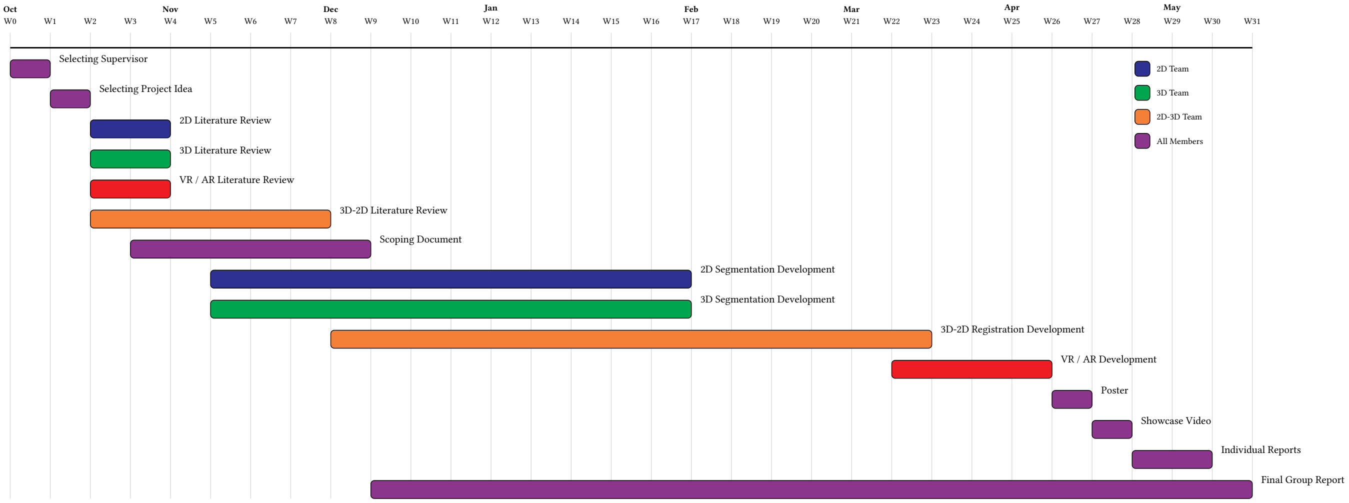

Fig. A1. A Gantt chart showing the initial project plan.

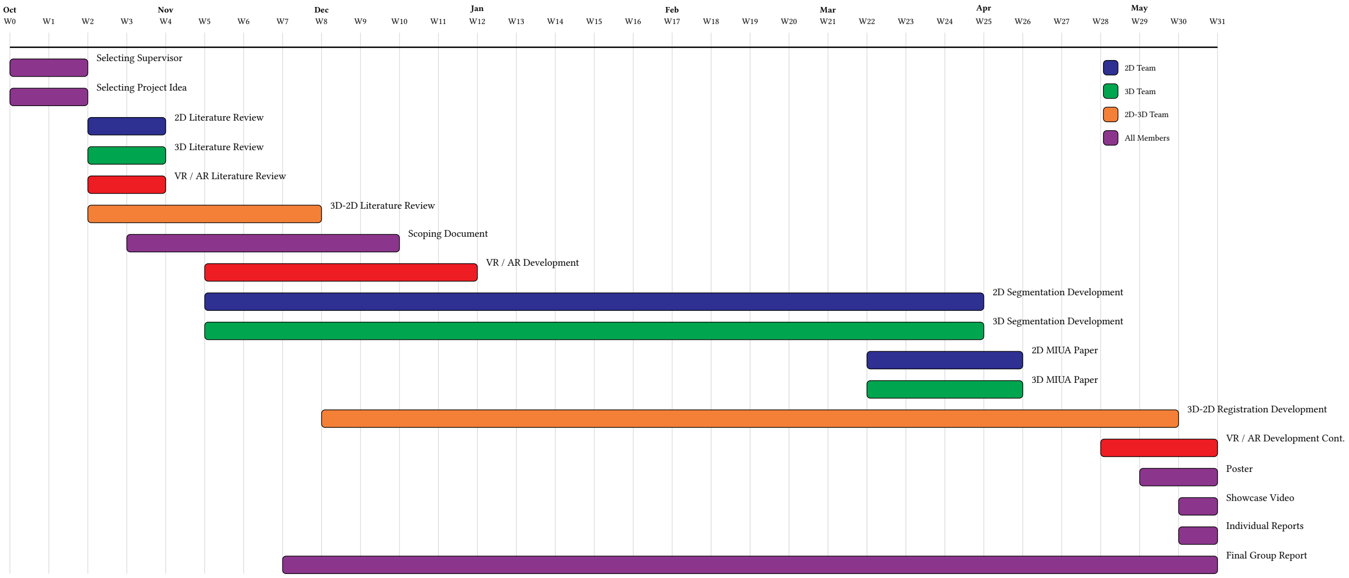

Fig. A2. A Gantt chart showing the final project timeline.

maintain momentum on the core deliverables, but we also succeeded in submitting two conference papers to MIUA.

B SELF-REFLECTION AND SOFT SKILLS ACQUIRED.

In conclusion, we believe the project to be an overall success. This is largely attributed to the completion of all tasks associated with the project, including the implementation of concrete methods by which we can perform 2D segmentation of images and 3D segmentation of liver models to form an automated end-to-end pipeline by which we can perform 3D-2D registration and have this solution presented on an augmented reality platform. Furthermore, when compared to other teams participating in the registration task (from the P2ILF challenge paper), our methods were able to outperform all six competing teams by a substantial margin in all three tasks, which demonstrate the practical successes of our solution. Whilst not an original aim set out at the beginning of the project, our work in the 2D segmentation and 3D segmentation tasks led to the creation of novel approaches which ultimately led to us submitting two papers to MIUA 2025, contributing to another success of the project.

We are satisfied with the overall success of the project, although in hindsight, we noticed a number of optimisations we could have made to our workflow that would have saved significant delays between tasks and ultimately could have avoided the busy period experienced before the final deadline. Difficulties arose due to 3D-2D registration's dependency on 2D and 3D segmentation tasks. Due to difficulties associated with 2D segmentation tasks, a group member was borrowed from the 3D-2D registration team to help. This was effective in finalising a solution to 2D segmentation, however, the delay was impactful and further passed on to aspects of 3D-2D registration due to the reduced workforce on this team during that time. Additionally, submitting two full conference papers to MIUA was considered an achievement for the group.

From a managerial perspective, our use of GitHub Enterprise (using an organisation) and version control was deemed an effective approach to working on the project - one that worked well for all members of the group. In this way, each group member, and the supervisor, was kept fully up-to-date on the latest developments from everyone in the project and new tasks were given out in response to these developments, ensuring that no one remained idle between meetings. This allowed the group to assess overall progress on the project at any given time, allowing for planning of next steps and the adjusting of project timings so that we could stay on track for set goals and the final deadline. This approach was amplified by our group's persistent and clear communication outside of these meetings, alongside work often performed in the presence of other group members.

As a group, we were able to learn and gain experience with a wide range of new soft skills associated with the project, including interpersonal skills, detailed communication and time management. Through consistent regular meetings with a more experienced academic and researcher where we demonstrated our progress towards a larger goal, we were able to experience a style of meeting akin to that commonly seen in the workplace where larger projects are concerned. This taught us to convey concepts in a way where anyone could understand, as separate teams were not entirely familiar with the works being researched in each task, thus developing our professional communication skills. By balancing work with this project against other university deadlines and exams, as well as two submissions to a conference, we were able to experience a fast paced, high workload. This constantly forced us to adapt to quick changes in direction. Through this, we were able to learn how to structure our time better and manage the multiple tasks we had at once, thus improving our skills in adaptability and time management. Finally, as the project was largely focused in areas in which none of us had prior experience, we were very much challenged in our abilities of problem solving, critical thinking and teamwork whilst theorising and attempting to solve our respective tasks. This further developed these soft skills acquired during the project, which will no doubt help each member of the group in our future ventures in the workforce.

C TABLES AND FIGURES

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

Table C1. P2ILF challenge 2D segmentation task results [Ali et al. 2025]. Teams are evaluated using our implementations of the following metrics. **P** is precision, **D** is DSC, and **G** is the symmetric distance metric in Equation 10. Results are in the order of ridge, falciform ligament, and silhouette. $\bar{\mathbf{P}}$, $\bar{\mathbf{D}}$, and $\bar{\mathbf{G}}$ are the mean across the three landmarks. The best mean results are highlighted in bold.

Ridge	Ligament	Silhouette	Loss (4_7)	Loss (4_11)	Loss (4_17)	Loss (11_2)	Loss (11_6)	Loss (11_7)	Mean
0.5	1	0.5	0.275	0.284	0.240	0.569	0.570	0.529	**0.411**
0.5	1	1	0.398	0.419	0.374	0.761	0.764	0.717	0.572
1	0.5	1	0.453	0.481	0.459	0.770	0.782	0.737	0.614
1	1	0.5	0.364	0.375	0.332	0.705	0.704	0.659	0.523
5	1	1	1.188	1.233	1.203	1.999	1.961	1.943	1.588
1	5	1	0.762	0.732	0.512	1.965	1.851	1.780	1.270
1	1	5	1.462	1.573	1.469	2.408	2.527	2.365	1.967
0.5	5	1	0.665	0.635	0.430	1.773	1.732	1.679	1.152
1	5	0.5	0.638	0.619	0.388	1.749	1.722	1.485	1.100

Table C2. Loss values for different combinations of ridge, ligament, and silhouette weight coefficients. The loss metrics are reported for multiple patient instances (4_7, 4_11, 4_17, 11_2, 11_6, 11_7), along with the mean loss. Lower loss values indicate better performance.