Neural Networks 178 (2024) 106469

Contents lists available at ScienceDirect

Neural Networks

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

Full Length Article

Towards safer robot-assisted surgery: A markerless augmented reality framework

Ziyang Chen [a], Laura Cruciani [a], Ke Fan [a,*], Matteo Fontana [b], Elena Lievore [b], Ottavio De Cobelli [b,c], Gennaro Musi [b,c], Giancarlo Ferrigno [a], Elena De Momi [a,b]

[a] Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milano, 20133, Italy
[b] European Institute of Oncology, Department of Urology, IRCCS, Milan, 20141, Italy
[c] University of Milan, Department of Oncology and Onco-haematology, Faculty of Medicine and Surgery, Milan, 20122, Italy

ARTICLE INFO

Keywords:
Robot-assisted surgery
Markerless augmented reality
Stereo reconstruction
Segmentation
Da Vinci research kit

ABSTRACT

Robot-assisted surgery is rapidly developing in the medical field, and the integration of augmented reality shows the potential to improve the operation performance of surgeons by providing more visual information. In this paper, we proposed a markerless augmented reality framework to enhance safety by avoiding intra-operative bleeding, which is a high risk caused by collision between surgical instruments and delicate blood vessels (arteries or veins). Advanced stereo reconstruction and segmentation networks are compared to find the best combination to reconstruct the intra-operative blood vessel in 3D space for registration with the pre-operative model, and the minimum distance detection between the instruments and the blood vessel is implemented. A robot-assisted lymphadenectomy is emulated on the da Vinci Research Kit in a dry lab, and ten human subjects perform this operation to explore the usability of the proposed framework. The result shows that the augmented reality framework can help the users to avoid the dangerous collision between the instruments and the delicate blood vessel while not introducing an extra load. It provides a flexible framework that integrates augmented reality into the medical robotic platform to enhance safety during surgery.

1. Introduction

Robot-Assisted Surgery (RAS) improves patient outcomes in both intra-operative operation and post-operative recovery compared to traditional open surgery, and it also provides the possibility to integrate artificial intelligence into this platform for autonomous and safe operation. An emerging technology, i.e., Augmented Reality (AR) fusing virtual targets on real scenes, provides more visual information to users and has been introduced in the field of robotic surgery to improve safety (Kim & Kim, 2023; Rodler et al., 2023; Su et al., 2020). Generally, the surgeon needs to acquire the pre-operative images of the patient, such as Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), for pre-operative planning. These CT/MRI slices can be segmented using software such as 3D Slicer to create the pre-operative 3D model and then projected onto the intra-operative images to implement the AR effect. These augmented intra-operative images can guide the operation of the surgeons by providing more visual information and improve the operation performance. One major challenge is how to accurately localize the intra-operative soft tissues or organs so that the pre-operative model can be registered with the intra-operative target, to implement the overlapping of the pre-operative model on the corresponding position of intra-operative images. In Bianchi et al. (2021), Schiavina et al. (2021), the authors implemented the AR in robot-assisted radical prostatectomy by overlapping the pre-operative 3D model on the endoscopic images, and they used the software vMIX (StudioCoast Pty Ltd, Australia) to manually align the position of the pre-operative model, which hinders the practice in real-time AR visualization during surgery. Similarly, a manual alignment between the pre-operative model and the intra-operative anatomy was used in some other operations, such as the robotic thyroidectomy (Lee et al., 2018), the transoral surgery (Chan et al., 2020) and the partial nephrectomy (Backer et al., 2023). In Wendler, van Leeuwen, Navab, and van Oosterom (2021), the authors introduced three possible solutions to implement AR registration including landmarks, laparoscopic video and intra-operative ultrasound to recognize the 3D position of surgical scenes, and they summarized that the laparoscopic video-based approach will be the mainstream since it does not require external hardware. Also, the authors in Qian, Wu, DiMaio, Navab, and Kazanzides (2019) introduced the registration process using the robotic instruments by pointing at pre-installed markers or the extra projector-camera system. To implement the automatic AR alignment

* Corresponding author.
E-mail address: ke.fan@polimi.it (K. Fan).

https://doi.org/10.1016/j.neunet.2024.106469
Received 28 March 2024; Received in revised form 1 June 2024; Accepted 16 June 2024
Available online 19 June 2024

using the laparoscopic video based on the da Vinci Surgical System (dVSS), the authors in Penza, De Momi, Enayati, Chupin, Ortiz, and Mattos (2017) proposed to track the interested soft tissue and then recover the corresponding 3D position information of the soft tissue using stereo reconstruction. The experimental result showed that the AR effect can be implemented on the intra-operative images. Nevertheless, this approach relied on an external device to manually draw the boundary of the target of interest at the beginning of the operation, and the pre-operative model was substituted using a simple ellipsoid that is different from the intra-operative tissue.

Many stereo reconstruction methods have been proposed to reconstruct the 3D scene by estimating the disparity map, and the depth information can be recovered by a transformation with the focal length and baseline of the stereo endoscope. Chang and Chen (2018) is a typical stereo reconstruction network, the authors adopted the Convolutional Neural Network (CNN) with embedding a Spatial Pyramid Pooling (SPP) module to extract the high-level features of the stereo image, and then they were concatenated as a 4D cost volume and fed into a stacked hourglass architecture for disparity estimation. To simplify this hourglass architecture, the authors in Guo, Yang, Yang, Wang, and Li (2019) adopted fewer skipping connections in the decoder and designed a novel 4D cost volume by calculating group-wise correlation. Next, Yang, Manela, Happold, and Ramanan (2019) was proposed to predict image pairs with a high resolution, different-level feature maps were employed to foster the multiple cost volumes and then gradually connected to estimate the disparity map based on a coarse-to-fine manner. More methods (Gu et al., 2020; Liu, Yu, & Qi, 2022; Shen, Dai, & Rao, 2021) were proposed to design different strategies to foster the cost volume since it is a key factor for the final prediction. Different from the traditional CNN architecture, attention-based transformer (Vaswani et al., 2017) provides a new network architecture and vision transformer (Dosovitskiy et al., 2020) opens the path to utilize this module in the vision field by encoding the image into many tokens. The works in Cheng et al. (2022), Li et al. (2021) started to fuse the transformer and CNN for the disparity estimation and their results showed a satisfactory performance by evaluating a public stereo endoscopic dataset (Allan et al., 2021).

3D reconstruction recovers the depth information of the whole intra-operative scene, which raises another issue, i.e., keeping the region of interest while removing other background points for an accurate registration between the pre-operative model and the intra-operative target. Hence, a segmentation neural network (Moccia, De Momi, El Hadji, & Mattos, 2018; Song, Wu, Song, Zhang & and Stojanovic, 2023) needs to be integrated to distinguish the background and the target by predicting a binary mask. UNet (Ronneberger, Fischer, & Brox, 2015) is the most representative model in medical image segmentation, and it adopted a U-shape architecture with downsampling and upsampling operation and implemented the fusion of different-level features using skipping connection. Following the similar UNet architecture, the authors in Zhuang (2018) proposed a multi-path feature fusion strategy by combining two UNet networks, the authors in Oktay et al. (2018) integrated an attention gate module to enhance the learning of target structures, and the authors in Zhou, Siddiquee, Tajbakhsh, and Liang (2019) designed a nested UNet architecture by densely aggregating different-scale features. The emerging transformer module has been utilized in the segmentation field. In Chen et al. (2021), Hatamizadeh et al. (2022), Zheng et al. (2021), the authors employed the transformers as the encoder to extract the features, and then adopted the CNN architecture as the decoder to predict the segmentation mask. More recently, a fancy model named Segment Anything (SAM) (Kirillov et al., 2023) was released which had been trained based on 11 million images and showed a strong generalization ability in various segmentation tasks. With the explosion of big data, it can be foreseen that such kind of models will lead a new era because of the promising zero-shot performance.

Intra-operative bleeding is a risky situation that affects surgical quality and post-operative recovery of patients, and it generally occurs due to unintentional collisions between surgical instruments and delicate blood vessels (arteries or veins) during surgery. Therefore, we propose a markerless augmented reality framework to mitigate the occurrence of this situation. Unlike existing approaches such as manual localization or using pre-installed markers, we combine the stereo reconstruction and segmentation networks to localize the intra-operative soft tissue. The proposed framework integrating the advanced neural networks can implement the visualization of the pre-operative model, and it also performs the minimum distance detection between the instruments and the blood vessel to avoid a dangerous collision. Furthermore, it does not rely on the additional external device, which means high generalization to other robotic systems and tasks. Overall, this work aims to solve an engineering problem of intra-operative bleeding in robotic surgery. It also has two scientific contributions: (a) It proposes a machine vision-based markerless localization solution to implement AR assistance; (b) It demonstrates the comprehensive performance of state-of-the-art stereo reconstruction and segmentation models, which can be a good reference for other researchers working on these models for their needs. The specific contributions can be listed as follows:

(1) A markerless augmented reality framework is proposed to visualize the pre-operative model on intra-operative scenes, and it provides the minimum distance detection between the surgical instruments and the delicate blood vessel for safety.

(2) A comprehensive evaluation of advanced neural networks in stereo reconstruction and segmentation fields is performed to find the best combination to accurately and fastly recover 3D information of the intra-operative blood vessel.

(3) A user study involving ten human subjects who performed a robot-assisted lymphadenectomy based on the da Vinci Research Kit in a dry lab is achieved to explore the usability of the proposed AR framework compared to the standard setup.

The remainder of this paper is structured as follows. Section 2 describes the details of our proposed AR framework. In Section 3, it presents the framework evaluation metrics and the specific experimental protocol of the usability study, and the results are given in Section 4. Section 5 discusses the findings and limitations of our work, and the conclusion of this paper and future work are drawn in Section 6.

2. Methodology

The proposed augmented reality framework is shown in Fig. 1, and it is integrated into the popular da Vinci Research Kit (dVRK, Intuitive Surgical Inc., US, and Johns Hopkins University). The stereo image pair is input into a stereo reconstruction network to estimate the left disparity map, and the left image is also input into a segmentation network to segment the blood vessel. The intra-operative blood vessel can then be generated in 3D space by combining the disparity map and the binary mask. The pre-operative model is used to perform registration with the intra-operative blood vessel so that it can be projected onto the corresponding position of the endoscopic image pairs to implement an AR effect. Furthermore, the minimum distances between the surgical instruments and the soft tissue are calculated based on the dVRK kinematics and the 3D position of the reconstructed blood vessel. The specific description of this framework is given below.

2.1. dVRK system and calibrations

The dVRK system is an open-source robotic platform consisting of the hardware of the first-generation da Vinci surgical system and customized software and electronics. It can be mainly divided into the leader side and the follower side. There are two Patient Side Manipulators (PSMs) that mount various surgical instruments such as the large needle driver at the follower side, and a stereo endoscope (1920×1080

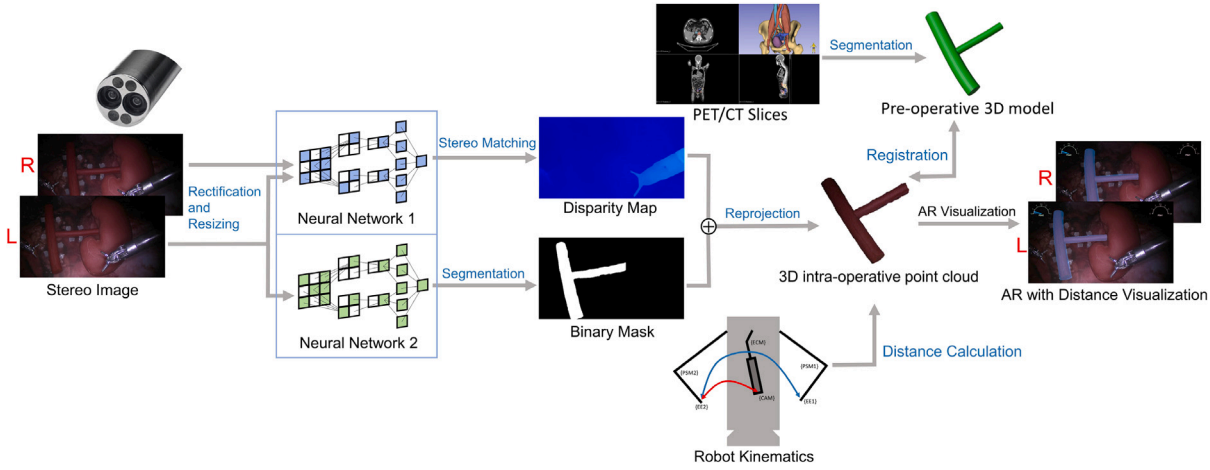

Fig. 1. The architecture of the markerless augmented reality framework. The image pair is fed into two networks to estimate the disparity map and the binary mask. Then, the disparity pixels that belong to the blood vessel are reprojected into 3D space to reconstruct the intra-operative blood vessel. The pre-operative model is overlapped on the corresponding region of the endoscopic images by registration, and the minimum distance between the instruments and the blood vessel is detected based on the dVRK kinematics and the 3D reconstructed blood vessel.

resolution) mounted on an Endoscopic Camera Manipulator (ECM) is used to capture in vivo surgical scenes. By adjusting the position of the Setup Joint (SUJ) connecting these robotic arms, the respective Remote Center of Motion (RCM) can be positioned at the skin entry point of the abdomen (Penza et al., 2017). On the other hand, the lead surgeon can view the endoscopic scenes through a High-Resolution Stereo Viewer (HRSV). This allows the surgeon to simultaneously view the left and right images captured by the stereo endoscope, thereby perceiving the depth of the intra-operative scenes. In addition, the surgeon can remotely control the movements of the PSMs and ECM by operating Master Tool Manipulators (MTMs) and a foot pedal tray (Chen et al., 2022; Col et al., 2020). The end posture of the surgical instrument is highly consistent with the end posture of the MTM, ensuring the high precision of teleoperated surgery.

Fig. 2 shows the details of our dVRK system, and it also presents the reference frame definition adopted in our framework. The Cartesian position of each instrument can be obtained from the dVRK kinematics, and a 9 × 6 chessboard with a square length of 1 cm is adopted for our calibrations. First, the camera calibration is done based on Zhang's calibration approach (Zhang, 2000) to generate the intrinsic and extrinsic parameters for the image rectification, undistortion and projection. Then, a hand-to-hand calibration is conducted to search for the rigid transformation between left and right end effectors based on Horn's method (Horn, 1987), since the position subscribed from the direct kinematics is not accurate enough. By collecting 40 non-collinear points, we can generate the rigid transformation matrix T_{EE1}^{EE2} and transfer the 3D points in *{EE1}* to *{EE2}*. Furthermore, a hand-eye calibration is performed to obtain the transformation between the left end effector *{EE2}* and the left camera *{L_CAM}*. We operate the left end effector to point at 54 corner points of the chessboard to obtain the 3D coordinates, and the corresponding 2D coordinates on the left image are obtained using cv.findChessboardCorners and cv.cornerSubPix (OpenCV) functions so that the transformation can be calculated based on the Random Sample Consensus (RANSAC) scheme (cv.solvePnPRansac function) (Bradski, 2000; Fischler & Bolles, 1981). Since our end effector is referencing the frame *{ECM}*, the transformation matrix $T_{ECM}^{L\_CAM}$ is generated based on the hand-eye calibration.

2.2. Intra-operative blood vessel reconstruction

The stereo image is first rectified to align the polar lines on the horizon axis, and the resolution is resized from 1920 × 1080 to 640 × 360 to speed up the framework. Then, the images are fed into a stereo reconstruction network to estimate the disparity map (Yang et al., 2019). In this way, we can reproject the disparity map to generate the 3D intra-operative point cloud. The conversion between the estimated disparity value $\hat{d}$ and the depth value $\tilde{d}$ is formulated as,

$$\tilde{d}(i,j) = \frac{bf}{\hat{d}(i,j)} \tag{1}$$

where i, j are the pixel position on the 2D image, b is the baseline of the stereo camera, and f is the focal length.

However, the reconstructed point cloud contains the whole intra-operative scene, and we need to extract the region of interest (the blood vessel) for registration. Considering that the estimated disparity map is referenced to the left rectified image, we propose to adopt another segmentation network to generate the binary mask of the region of interest. The left rectified image is fed into the segmentation network to estimate the indices of the blood vessel and the background. Next, we use the disparity pixels that belong to the blood vessel by referencing these indices and reproject them into 3D space to reconstruct the intra-operative blood vessel. The 3D position of the reconstructed blood vessel directly influences the quality of the following registration and the distance detection, so a comprehensive comparison study including the stereo reconstruction and segmentation networks is provided in Section 3 to determine the best combination in our framework. Finally, two post-processing approaches are adopted to improve the segmentation estimation, including mask boundary eroding and small object removal. Mask boundary eroding can remove the possible misclassified pixels near the boundary of the blood vessel, and the small object removal is used to remove the possible outliers in other regions, which can refine the segmentation quality in some challenging cases such as the blurred scenes caused by the fast movement of the instruments.

The reconstructed 3D blood vessel is referenced to the rectified left camera coordinate system *{Rec_L_CAM}*, and we transform it to the reference *{ECM}* for the following distance calculation based on the equation,

$$P_{intra\_obj}^{ECM} = \left(T_{ECM}^{L\_CAM}\right)^{-1} * \left(T_{L\_CAM}^{Rec\_L\_CAM}\right)^{-1} * P_{intra\_obj}^{Rec\_L\_CAM} \tag{2}$$

where $P_{intra\_obj}^{Rec\_L\_CAM}$ is the reconstructed 3D points of the intra-operative blood vessel referenced to the rectified left camera coordinate system, $T_{L\_CAM}^{Rec\_L\_CAM}$ is the transformation between the unrectified left camera system and the rectified left camera system obtained from the stereo image rectification, and $T_{ECM}^{L\_CAM}$ is obtained based on the hand-eye calibration.

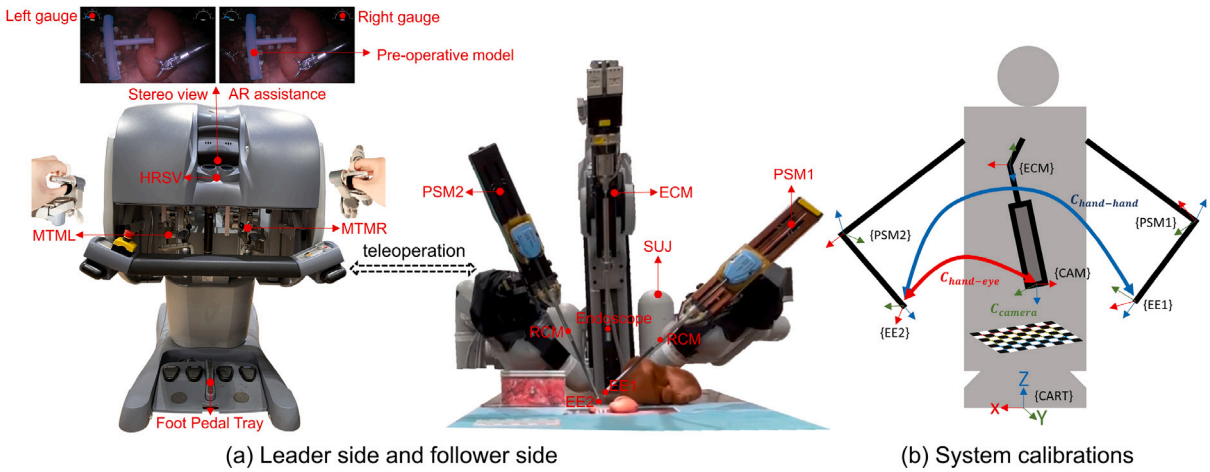

Fig. 2. The presentation of the dVRK system in a dry lab. In (a), the user can operate the MTMs and observe the surgical scenes using the HRSV at the leader side, and the surgical instruments mounted on the PSMs will perform the operation following the remote control of the user at the follower side. In (b), it shows the three types of system calibrations, including camera calibration, hand-hand calibration and hand-eye calibration.

2.3. Registration between pre-operative and intra-operative targets

The 3D pre-operative model is generally created using CT or MRI, and then software such as 3D Slicer is used to segment the region of interest to generate the 3D structure. In our case, we use 3D modeling software, Blender, to create a 3D pre-operative model for simplicity, since we do not have an external device to capture these CT/MRI slices. Next, we can perform registration between the pre-operative model and the intra-operative reconstructed point cloud. Here, the pre-operative blood vessel is a mesh model while the intra-operative one is a point cloud, so we sample plenty of points from the mesh model for the registration process. Considering that there is an apparent position difference between the pre-operative model and the intra-operative blood vessel in the initial state, the global registration-based RANSAC algorithm (Fischler & Bolles, 1981) is adopted to conduct the initial transformation, and it is only implemented once at the beginning. Then, the local registration-based Iterative Closest Point (ICP) (Besl & McKay, 1992) algorithm is performed to finetune the position of the pre-operative model. It is found that this registration strategy can accurately and fastly register the models in our experiment.

2.4. Distance detection and AR visualization

Our framework not only provides the augmented pre-operative model visualization on intra-operative images, but also detects the minimum distances between the surgical instruments and the blood vessel. After aligning the position of the pre-operative model with the intra-operative blood vessel, we could overlap the pre-operative model on the left intra-operative scenes by the transformation $P^{L\_img}_{pre\_obj}$,

$$P^{L\_img}_{pre\_obj} = K_L * T^{L\_CAM}_{ECM} * T^{ECM}_{BL} * P^{BL}_{pre\_obj} \tag{3}$$

where $P^{BL}_{pre\_obj}$ is the 3D pre-operative points referenced to the Blender coordinate system $\{BL\}$, T^{ECM}_{BL} is the transformation matrix generated by the registration, $T^{L\_CAM}_{ECM}$ is the matrix obtained from the hand-eye calibration, and K_L contains the intrinsic and distortion matrices of the left camera obtained from the camera calibration. Similarly, we can use the following equation to project the pre-operative model on the right intra-operative scenes $P^{R\_img}_{pre\_obj}$,

$$P^{R\_img}_{pre\_obj} = K_R * T^{R\_CAM}_{L\_CAM} * T^{L\_CAM}_{ECM} * T^{ECM}_{BL} * P^{BL}_{pre\_obj} \tag{4}$$

where $T^{R\_CAM}_{L\_CAM}$ is the transformation between the left and right camera coordination systems, and K_R contains the intrinsic and distortion matrices of the right camera. Then, we can observe that the pre-operative model is overlapped on the corresponding regions of the left and right images, respectively.

Finally, we calculate the minimum distance between the surface of the instruments and the reconstructed blood vessel based on the fast k-nearest-neighbor search strategy (Garcia, Debreuve, & Barlaud, 2008; Williams, 2022). The Cartesian position of the end effectors and the RCM points of instruments are subscribed from the robot kinematics so that we can model the instruments as cylinders with a radius of 4 mm and sample them as the point clouds (here, the position of PSM1 is aligned to PSM2 by the hand-hand rigid transformation T^{EE2}_{EE1}). Two gauges are provided in the left upper and right upper corners of intra-operative images, and they can visualize the respective minimum distance of left and right instruments. Also, the color of the pre-operative model automatically changes according to the smaller distance by comparing the left and right minimum distances to remind surgeons during operation.

3. Experimental protocol and performance metrics

3.1. Framework characterization evaluation

3.1.1. Reconstruction and segmentation networks

The prerequisite for performing AR visualization and distance detection is that the intra-operative position of the soft tissue needs to be accurately restored, so we explored 14 state-of-the-art methods in the stereo reconstruction field to find the best model that can be utilized in the medical scenes. Among them, ELAS (Geiger, Roser, & Urtasun, 2011) is an optimation based method while others (Chang & Chen, 2018; Cheng et al., 2020, 2022; Garg et al., 2020; Gu et al., 2020; Guo et al., 2019; Li et al., 2021, 2022; Liu et al., 2022; Shen et al., 2021; Xu, Cheng, Guo, & Yang, 2022; Xu, Wang, Ding, & Yang, 2023; Yang et al., 2019) utilize the neural networks. The stereo endoscopic dataset, SERV-CT (Edwards, Psychogyios, Speidel, Maier-Hein, & Stoyanov, 2022), was adopted to conduct the quantitative evaluation of these methods, and it contains sixteen image pairs captured from porcine samples based on the dVSS and provides the dense ground truth. To reconstruct the endoscopic scenes, we run these models based on their official weights without any task-specific fine-tuning for generalization (Lu, Jayakumari, Richter, Li, & Yip, 2021). A set of accuracy-related metrics were chosen to evaluate the reconstruction error by comparing the estimated depth with the provided ground truth, both expressed in millimeters. The metrics include Median Absolute Error (MeAE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Absolute

Relative Error (Abs Rel), Squared Relative Error (Sq Rel), as well as δ_{ratio} (Eigen, Puhrsch, & Fergus, 2014; Zhao, Sun, Zhang, Tang, & Qian, 2020).

$$\text{MeAE} = \text{Median}\left\{\tilde{d}(i,j) \in S \mid |\tilde{d}(i,j) - d'(i,j)|\right\} \tag{5}$$

$$\text{MAE} = \frac{1}{|S|}\sum_{(i,j)}|\tilde{d}(i,j) - d'(i,j)| \tag{6}$$

$$\text{RMSE} = \sqrt{\frac{1}{|S|}\sum_{(i,j)}|\tilde{d}(i,j) - d'(i,j)|^2} \tag{7}$$

$$\text{Abs Rel} = \frac{1}{|S|}\sum_{(i,j)}\frac{|\tilde{d}(i,j) - d'(i,j)|}{d'(i,j)} \tag{8}$$

$$\text{Sq Rel} = \frac{1}{|S|}\sum_{(i,j)}\frac{|\tilde{d}(i,j) - d'(i,j)|^2}{d'(i,j)} \tag{9}$$

$$\delta_{ratio} : \%\ \text{of}\ \tilde{d}(i,j)\ \text{s.t.}\ \max\left(\frac{\tilde{d}(i,j)}{d'(i,j)}, \frac{d'(i,j)}{\tilde{d}(i,j)}\right) < \tau \tag{10}$$

where S is the set of predicted depth values for each frame, $\tilde{d}(i,j)$ is the predicted depth value related to pixel in position (i,j) and $d'(i,j)$ is the ground truth of depth value. The last metric evaluates the depth fluctuation error between the reconstructed points and the ground truth, and three different thresholds $\tau \in [1.25^1, 1.25^2, 1.25^3]$ were adopted. Unlike the other metrics, the higher δ_{ratio} means the better reconstruction result. Also, we provided the single-frame inference time when evaluating these models.

Another segmentation network is required to extract the soft tissue of interest, so 8 state-of-the-art segmentation methods (Chen et al., 2021; Hatamizadeh et al., 2022; Kirillov et al., 2023; Oktay et al., 2018; Ronneberger et al., 2015; Zheng et al., 2021; Zhou et al., 2019; Zhuang, 2018) were evaluated for the blood vessel segmentation. The recent segmentation network SAM (Kirillov et al., 2023) has strong generalization in different fields, while other neural networks need to be trained. Hence, we captured six endoscopic videos containing the 3D-printed blood vessel based on the dVRK platform in a dry lab and extracted around 100 images from each video for the manual annotation (551 frames in total). We performed the annotation using Computer Vision Annotation Tool (CVAT) (Sekachev et al., 2020). Six-fold cross validation was adopted to train and evaluate the models. During the training process, we cropped the images into 128 × 128 patches and trained the models for 100 epochs based on a batch size of 64 (100000 patches in every epoch and 10% of images from the training images were used for the model validation). Two methods (Chen et al., 2021; Zheng et al., 2021) loaded the pre-trained weight following the authors' original configuration while other models were trained from the scratch. The images were also split into patches with the resolution of 128 × 128 without overlapping during the test phase. Common evaluation metrics in the segmentation field were used for a comprehensive comparison, including Dice coefficient ($\frac{2TP}{2TP+FP+FN}$), accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$), specificity ($\frac{TN}{TN+FP}$), sensitivity ($\frac{TP}{TP+FN}$), precision($\frac{TP}{TP+FP}$), the area of the Precision-Recall (PR) curve (Venugopal et al., 2022) as well as the inference time.

3.1.2. Other characterization

The quantitative performance of this framework needs to be measured for practicality, and the experimental platform is based on an NVIDIA RTX 3080 GPU in a local laptop. On the one hand, we provided the computational time of the proposed framework by calculating the time spent in 100 frames of dynamic scenes, and we also gave the specific time distribution in each component of this framework. On the other hand, we calculated the error distribution in different components of this framework. The system calibrations based on the dVRK introduce errors, and the respective performance metric can be formulated as:

• Camera calibration error E_{cam} considers the difference between the coordinates of reprojected 3D points on the 2D plane and the actual coordinates of the chessboard corners by calculating the root mean square value,

$$E_{\text{cam}} = \sqrt{\frac{1}{N_{\text{cam}}}\sum_{i=1}^{N_{\text{cam}}}\left\|\varepsilon_i^{\text{pro}} - \varepsilon_i^{\text{act}}\right\|_2^2} \tag{11}$$

where $\varepsilon_i^{\text{pro}}$ is the 2D points reprojected using the camera calibration parameters, while $\varepsilon_i^{\text{act}}$ denotes the actual 2D coordinates on the chessboard corners. N_{cam} is the number of the points used for the camera calibration, and $\|,\|_2$ represents the Euclidean norm.

• Hand-hand calibration error $E_{\text{hand\_hand}}$ was calculated based on the Cartesian position difference between the 3D points subscribed from the left end effector and the same points transformed from the right end effector based on the hand-to-hand matrix T_{EE1}^{EE2},

$$E_{\text{hand\_hand}} = \sqrt{\frac{1}{N_{\text{hh}}}\sum_{i=1}^{N_{\text{hh}}}\left\|\rho_i^L - T_{EE1}^{EE2}\rho_i^R\right\|_2^2} \tag{12}$$

where ρ_i^L and ρ_i^R are the 3D Cartesian positions of the left and right end effectors, respectively. N_{hh} is the number of the 3D points used for the hand-hand calibration.

• Hand-eye calibration error $E_{\text{hand\_eye}}$ was evaluated based on the pixel position difference between the reprojected 2D coordinates γ_i^{pro} from the 3D points of the left end effector using the hand-eye transformation matrix and the actual 2D coordinates γ_i^{act},

$$E_{\text{hand\_eye}} = \sqrt{\frac{1}{N_{\text{he}}}\sum_{i=1}^{N_{\text{he}}}\left\|\gamma_i^{\text{pro}} - \gamma_i^{\text{act}}\right\|_2^2} \tag{13}$$

where N_{he} is the number of the points adopted for the hand-eye calibration.

Furthermore, the errors of the reconstruction and segmentation networks have been reported in the last subsection, and the registration process also introduces errors. We calculated the Cartesian position error E_{regis} between the pre-operative blood vessel after registration using the transformation matrix T_{BL}^{ECM} and the reconstructed blood vessel,

$$E_{\text{regis}} = \sqrt{\frac{1}{N_{\text{re}}}\sum_{i=1}^{N_{\text{re}}}\left\|T_{BL}^{ECM}\varphi_i^{\text{pre-op}} - \varphi_i^{\text{recon}}\right\|_2^2} \tag{14}$$

where $\varphi_i^{\text{pre-op}}$ is the pre-operative 3D points, while φ_i^{recon} is the intra-operative 3D reconstructed points referenced to the frame *{ECM}*. N_{re} is the point number for the registration.

3.2. Framework usability study

To explore the usability of our proposed framework, a common surgical operation named lymphadenectomy was designed in a dry lab environment based on the proposal of an oncology surgeon. As shown in Fig. 3, the 3D-printed soft blood vessel (renal artery) and kidney were adopted to emulate the surgical scene, and the defined task is to remove the ten lymph nodes one by one (the white soft objects) near the blood vessel while not touching it. Frame 1 shows the initial position of the instruments, then the users operate the left instrument to clamp the left six lymph nodes and put them at the bottom left corner as shown in Frames 2 and 3, next the users operate the right instrument to remove the right four lymph nodes and put them at the same corner as shown in Frames 4 and 5, finally the users move the instruments to the initial position in Frame 6. Frame 3 and Frame 5 are challenging cases since the instruments are more likely to collide with the blood vessel. Ten human subjects (6 males and 4 females) who have a biomedical background were invited to join our experiment. Before the experiment, the users had spent 10 min for them to become familiar with the dVRK system. Then, the experiment was repeated in three rounds and we analyzed the user data of the last round to avoid the possible influence of the learning curve. In each round, two different modalities were

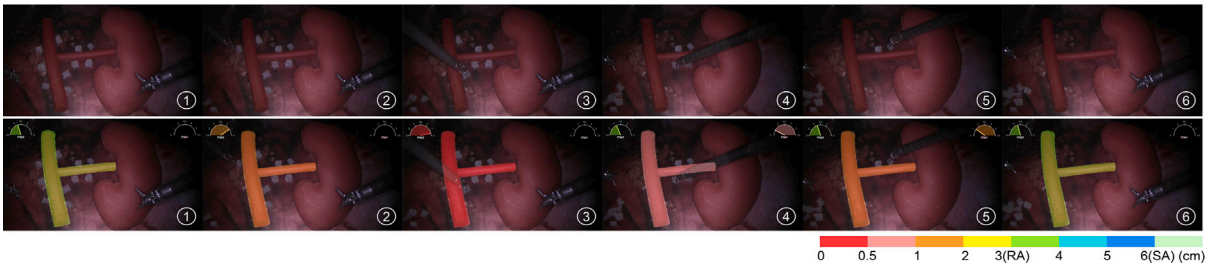

Fig. 3. An emulated lymphadenectomy based on the dVRK platform in a dry lab. The first row presents the standard endoscopic scenes, while the second row shows the augmented scenes with AR assistance. The two gauges show the respective minimum distance between the two instruments and the blood vessel. The pre-operative model is overlapped on the intra-operative blood vessel, and its color automatically changes according to the smaller distance by comparing the left and right distances. The defined task is to remove the ten lymph nodes without touching the delicate blood vessel for safety.

Table 1
Quantitative depth estimation result based on the SERV-CT stereo endoscopic dataset (The image resolution is 720 × 576).

	MeAE (mm)	MAE (mm)	RMSE (mm)	Abs Rel	Sq Rel	$\delta < 1.25^1$	$\delta < 1.25^2$	$\delta < 1.25^3$	Time (s)
HybridStereoNet (Cheng et al., 2022)	46.98 ± 19.32	54.30 ± 16.86	73.66 ± 13.79	0.67 ± 0.08	70.27 ± 18.33	0.14 ± 0.07	0.24 ± 0.12	0.39 ± 0.21	0.52 ± 0.01
ACVNet (Xu et al., 2022)	8.43 ± 13.46	17.88 ± 12.24	28.26 ± 14.52	0.20 ± 0.12	10.12 ± 8.65	0.69 ± 0.19	0.72 ± 0.19	0.77 ± 0.20	0.42 ± 0.01
STTR (Li et al., 2021)	3.82 ± 6.81	15.65 ± 10.24	42.26 ± 22.28	0.22 ± 0.20	31.88 ± 33.58	0.83 ± 0.18	0.88 ± 0.14	0.94 ± 0.06	0.43 ± 0.00
ELAS (Geiger et al., 2011)	3.81 ± 3.14	9.05 ± 8.51	18.86 ± 19.72	0.13 ± 0.16	11.73 ± 29.87	0.87 ± 0.18	0.95 ± 0.07	0.98 ± 0.04	0.05 ± 0.01
GraftNet (Liu et al., 2022)	1.27 ± 0.55	8.23 ± 4.59	35.42 ± 19.45	0.11 ± 0.08	23.43 ± 21.61	0.96 ± 0.03	0.96 ± 0.03	0.97 ± 0.02	0.33 ± 0.00
Cascade-Stereo (Gu et al., 2020)	1.80 ± 0.67	5.53 ± 3.00	13.15 ± 6.72	0.06 ± 0.02	2.15 ± 1.72	0.93 ± 0.04	0.96 ± 0.03	0.98 ± 0.02	0.34 ± 0.00
PSMNet (Chang & Chen, 2018)	1.43 ± 0.61	3.43 ± 1.94	7.13 ± 4.81	0.04 ± 0.02	0.72 ± 1.05	0.97 ± 0.03	0.98 ± 0.02	0.99 ± 0.02	0.40 ± 0.01
GwcNet (Guo et al., 2019)	1.77 ± 0.56	3.06 ± 1.25	5.22 ± 2.77	0.04 ± 0.01	0.35 ± 0.39	0.98 ± 0.03	1.00 ± 0.01	1.00 ± 0.00	0.31 ± 0.01
CFNet (Shen et al., 2021)	1.29 ± 0.49	3.04 ± 2.13	6.99 ± 7.21	0.04 ± 0.05	2.01 ± 6.16	0.98 ± 0.02	0.99 ± 0.02	1.00 ± 0.01	0.29 ± 0.00
W-Stereo-Disp (Garg et al., 2020)	1.35 ± 0.41	3.02 ± 1.38	6.59 ± 4.35	0.04 ± 0.02	0.71 ± 1.04	0.98 ± 0.03	0.99 ± 0.02	1.00 ± 0.01	0.42 ± 0.00
LEAStereo (Cheng et al., 2020)	1.44 ± 0.82	2.96 ± 1.28	6.29 ± 2.60	0.04 ± 0.01	0.58 ± 0.44	0.98 ± 0.02	1.00 ± 0.01	1.00 ± 0.00	0.52 ± 0.01
IGEV-Stereo (Xu et al., 2023)	1.12 ± 0.47	2.59 ± 1.18	5.64 ± 3.10	0.03 ± 0.01	0.40 ± 0.36	0.98 ± 0.02	1.00 ± 0.01	1.00 ± 0.00	0.32 ± 0.00
CREStereo (Li et al., 2022)	1.32 ± 0.71	2.38 ± 1.58	4.18 ± 2.90	0.03 ± 0.01	0.23 ± 0.27	0.99 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.14 ± 0.01
HSM (Yang et al., 2019)	1.10 ± 0.60	2.05 ± 1.17	3.72 ± 2.08	0.02 ± 0.01	0.17 ± 0.17	0.99 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	0.03 ± 0.00

performed in a random sequence for each user: Control (the standard endoscopic scene without the AR assistance), and Experiment (the endoscopic scene with the AR assistance).

Our AR framework not only visualizes the corresponding pre-operative model on the intra-operative blood vessel, but it also provides the visualization of the minimum distances between the instruments and the blood vessel. We added two gauges on the scenes to visualize the minimum distances (SA: 6 cm means the safe area and RA: 3 cm means the risk area), and the color of the pre-operative model changes according to the smaller distance between the left and right distances. Five performance metrics were utilized to observe if the AR framework can help improve the surgical performance,

• Minimum distance $D_{\min}$ between the instruments and the blood vessel:

$$D_{\min} = \min\left\{d_{1L}, d_{2L}, \ldots, d_{ML}, d_{1R}, d_{2R}, \ldots, d_{MR}\right\} \tag{15}$$

where d_{ML} is the minimum distance of the left instrument in the M-th frame, while d_{MR} denotes the minimum distance of the right instrument in the M-th frame during operation.

• Mean distance D_{mean} when the instruments are in the risk area of 3 cm:

$$D_{\text{mean}} = \frac{1}{\widetilde{M} + \widetilde{N}}\left(\sum_{m=1}^{\widetilde{M}} d_{mL} + \sum_{n=1}^{\widetilde{N}} d_{nR}\right) \tag{16}$$

where $\widetilde{M}$ is the number of the minimum distance points of the left instrument when they are smaller than 3 cm, while $\widetilde{N}$ represents the point number when the minimum distances of the right instrument are smaller than 3 cm.

• Collision number N_{c} when the distance points are smaller than the threshold r:

$$N_{\text{c}} = \sum_{m=1}^{M}\left\{1 \mid d_{mL} < r\right\} + \sum_{m=1}^{M}\left\{1 \mid d_{mR} < r\right\} \tag{17}$$

where r is defined as 0.5 cm in our case. Here, we regard it as one time of collision if the points remain smaller than 0.5 cm for one consecutive second during the task.

• Overall movement path S_{p} of the instruments during the operation:

$$S_p = \sum_{m=2}^{M}\left(\left\|C_L^m - C_L^{m-1}\right\|_2 + \left\|C_R^m - C_R^{m-1}\right\|_2\right) \tag{18}$$

where C_L^m is the 3D Cartesian coordinate of the left end effector in the mth frame, while C_R^m is the 3D coordinate of the right one in the mth frame.

• Execution time T_{exe} to perform the complete operation:

$$T_{\text{exe}} = T_{\text{end}} - T_{\text{start}} \tag{19}$$

where T_{start} and T_{end} are the start time and the end time of the task, respectively.

To investigate the feasibility and friendliness of the AR framework, the users were invited to fill out a System Usability Scale (SUS) questionnaire containing ten typical questions (Brooke et al., 1996) after the experiment. They evaluated the two systems (the standard system and the extended system with the AR assistance) by giving a score for each question (from score 1: strongly disagreement to score 5: strongly agreement), and the final SUS score of each user can be calculated as,

$$\text{SUS}_{\text{score}} = \left(\sum_{k=1,3,5,7,9}\left(S_k - 1\right) + \sum_{k=2,4,6,8,10}\left(5 - S_k\right)\right) * 2.5 \tag{20}$$

where S_k denotes the score of the kth question provided by the user.

4. Results

4.1. Results on framework characterization evaluation

4.1.1. Reconstruction and segmentation networks

Table 1 gives the quantitative evaluation result using these advanced stereo reconstruction models. The model HSM (Yang et al.,

Table 2
Quantitative segmentation result using the self-made dataset (resolution: 1920 × 1080) based on 6-fold cross validation.

	DSC	Accuracy	Specificity	Sensitivity	Precision	Area of PR Curve	Time (s)
Attention UNet (Oktay et al., 2018)	0.6708 ± 0.2296	0.9089 ± 0.0778	0.9059 ± 0.0856	0.9486 ± 0.0404	0.5860 ± 0.3075	0.7974 ± 0.1324	0.3509 ± 0.0018
Segment Anything (Kirillov et al., 2023)	0.9661 ± 0.0372	0.9957 ± 0.0041	0.9997 ± 0.0003	0.9404 ± 0.0633	0.9958 ± 0.0049	N/A	0.6219 ± 0.0071
UNETR (Hatamizadeh et al., 2022)	0.9780 ± 0.0082	0.9971 ± 0.0010	0.9986 ± 0.0006	0.9757 ± 0.0116	0.9805 ± 0.0093	0.9931 ± 0.0063	0.2056 ± 0.0014
SETR (Zheng et al., 2021)	0.9794 ± 0.0076	0.9973 ± 0.0009	0.9989 ± 0.0004	0.9747 ± 0.0126	0.9843 ± 0.0064	0.9933 ± 0.0068	0.1970 ± 0.0010
TransUNet (Chen et al., 2021)	0.9823 ± 0.0059	0.9976 ± 0.0006	0.9988 ± 0.0006	0.9818 ± 0.0101	0.9829 ± 0.0085	0.9985 ± 0.0015	0.3066 ± 0.0029
UNet++ (Zhou et al., 2019)	0.9825 ± 0.0062	0.9976 ± 0.0007	0.9986 ± 0.0007	0.9842 ± 0.0080	0.9809 ± 0.0099	0.9984 ± 0.0017	0.3498 ± 0.0016
LadderNet (Zhuang, 2018)	0.9835 ± 0.0061	0.9978 ± 0.0007	0.9989 ± 0.0005	0.9824 ± 0.0100	0.9846 ± 0.0071	0.9987 ± 0.0017	0.0898 ± 0.0012
UNet (Ronneberger et al., 2015)	0.9855 ± 0.0038	0.9980 ± 0.0005	0.9990 ± 0.0004	0.9842 ± 0.0072	0.9869 ± 0.0059	0.9990 ± 0.0010	0.3178 ± 0.0017

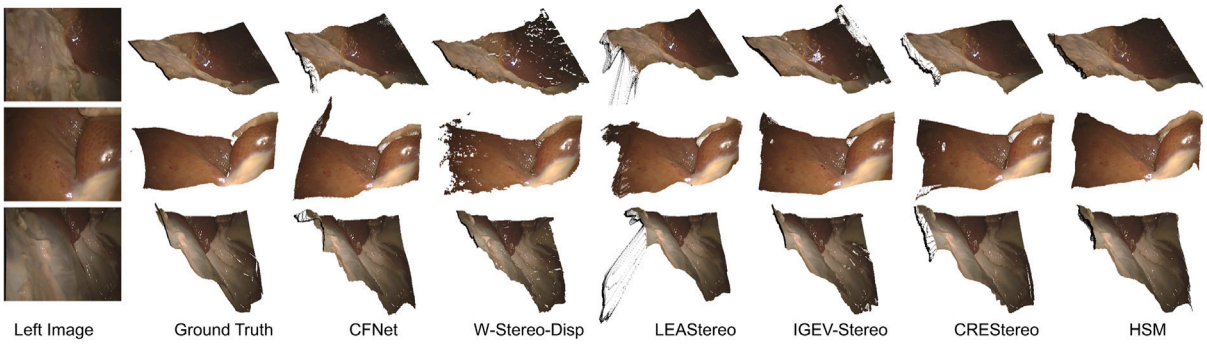

Fig. 4. Qualitative surgical scene reconstruction result. The reconstructed 3D surgical surfaces based on six representative models are provided.

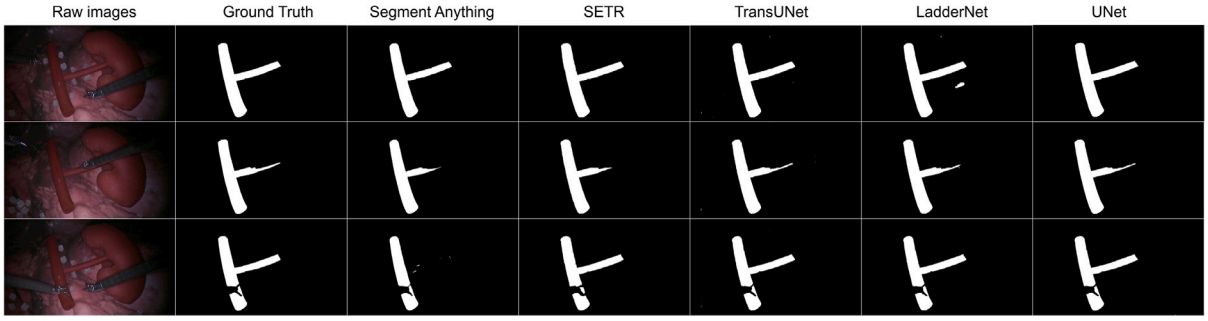

Fig. 5. Qualitative segmentation result using a self-made dataset containing the 3D-printed blood vessel captured from the dVRK platform.

Table 3
The computational time distribution of the framework (the image resolution is resized into 640 × 360).

Phase	Time (s)
Stereo image preprocessing	0.0095 ± 0.0017
Disparity map estimation	0.0365 ± 0.0018
Binary mask estimation with postprocessing	0.0470 ± 0.0015
Point cloud generation and alignment	0.0032 ± 0.0004
Distance calculation	0.0158 ± 0.0020
Registration between pre-op and intra-op targets	0.0015 ± 0.0002
Augmented reality visualization	0.0312 ± 0.0043
Whole pipeline	0.1448 ± 0.0079

Table 4
The error distribution in the framework.

Component	Metric	Value/Unit
Camera calibration	E_{cam}	0.60 ± 0.17 (pixels)
Hand-hand calibration	E_{hand_hand}	0.10 ± 0.05 (cm)
Hand-eye calibration	E_{hand_eye}	1.64 ± 0.80 (pixels)
Reconstruction	MAE	0.205 ± 0.117 (cm)
Segmentation	Dice	0.9855 ± 0.0038
Registration	E_{regis}	0.4328 ± 0.0385 (cm)

2019) gets the best performance in both accuracy and inference time. The qualitative comparison based on three frames is also provided in Fig. 4. HSM can reconstruct smoother soft tissue surfaces with fewer outliers than other methods, so we chose this model for depth estimation in our framework.

The quantitative segmentation result is provided in Table 2, and the qualitative comparison result is shown in Fig. 5. Here, we provided the specific boxes as the prompt for the mask estimation of the SAM model, and the area of the PR curve is not applicable to this model since it directly predicts the pixel classification instead of the probability. We noticed that UNet could provide a reliable segmentation quality in our phantom-based environment, so we adopted this model to estimate the mask in our experiment.

4.1.2. Other characterization

Table 3 presents the computational time distribution of this framework. Among them, stereo image preprocessing includes image subscribing, rectification and resizing from 1920 × 1080 to 640 × 360 based on a fixed scaling factor. The three phases, i.e., disparity map estimation, binary mask estimation and AR visualization, consume the most time in this framework, and the total time to process one frame is 0.1448 ± 0.0079s (6.91FPS), which can provide a smooth visual feedback during operation. Table 4 shows the error distribution in this framework. It can be noticed that these errors are small enough in our experiment.

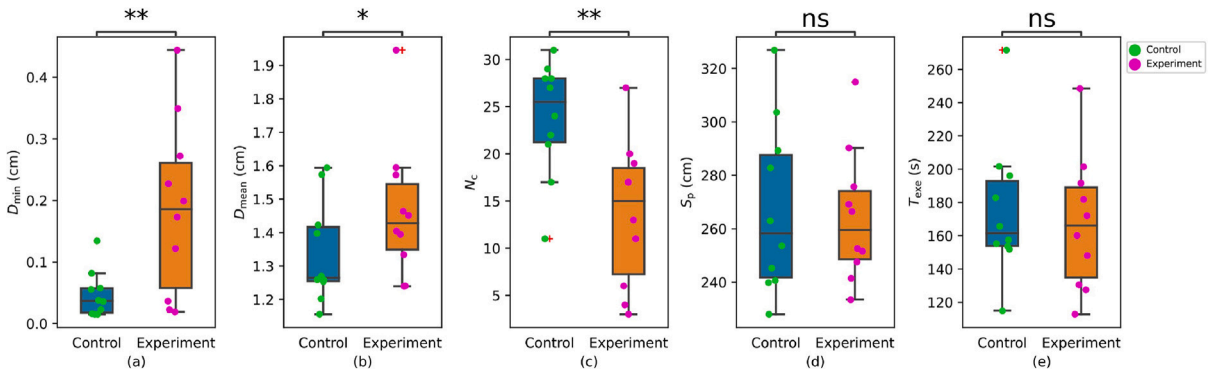

Fig. 6. The data of the ten human subjects based on five performance metrics. "Control" means the users complete the operation based on the standard endoscopic scenes, while "Experiment" is based on the scenes with the AR assistance. The result of the Wilcoxon signed-rank test is shown as ns : $0.05 < p \leq 1$, $*$: $0.01 < p \leq 0.05$, $**$: $0.001 < p \leq 0.01$, $***$: $0.0001 < p \leq 0.001$, and $****$: $p \leq 0.0001$.

Table 5
The average values and p values based on the users' data.

	Control	Experiment	P value
Minimum Distance D_{min} (cm)	0.0472	0.1864	0.0098
Mean Distance D_{mean} (cm)	1.3387	1.4641	0.0371
Collision Number N_c	23.8	13.7	0.0077
Overall Path S_p (cm)	267.28	264.32	0.5566
Execution Time T_{exe} (s)	175.11	167.44	0.2324

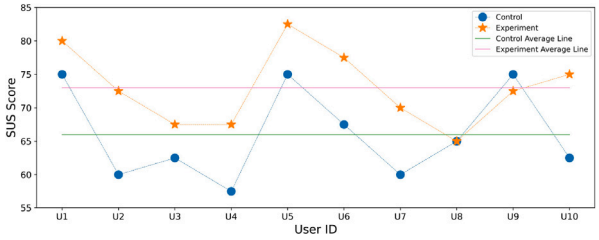

Fig. 7. The specific SUS score distribution provided by the users and the average SUS scores in two different modalities.

4.2. Results on framework usability study

Fig. 6 presents the box plots of the user data, and the Wilcoxon signed-rank test (p<0.05) is conducted to explore if there is a significant difference between the control modality and the experiment modality. It shows that there are significant differences in the minimum distance D_{min}, the mean distance D_{mean} and the collision count N_c. Table 5 also provides the average values using the five performance metrics. With the AR assistance, the minimum distance D_{min} increases from 0.0472 cm to 0.1864 cm, and the mean distance D_{mean} in the risk area also increases from 1.3387 cm to 1.4641 cm. When considering the collision number N_c, the value reduces from 23.8 to 13.7. The statistical differences show that the AR assistance reduces the collision probability between the instruments and the blood vessel during operation. Furthermore, there is no statistical difference in terms of overall path S_p and execution time T_{exe} , which indicates that the AR assistance does not introduce an extra load in operating the robot.

The final result of the SUS questionnaire is given in Fig. 7. The average SUS score of the control modality is 66, while the experiment modality has a higher SUS score of 73. The statistical test shows that there is a significant difference between the two modalities (p = 0.0104).

5. Discussion

AR is a popular direction in various fields, including robotic surgery, as it offers the possibility of increasing safety during operation, such as the distance visualization in our case or the visualization of some invisible tissues. However, an unsolved challenge is to localize the region of interest so that the pre-operative model can be registered with the corresponding intra-operative soft tissues or organs. In this work, we proposed a vision-based markerless localization approach to perform the AR assistance with the distance visualization on the intra-operative scenes, which releases the burden of the manual alignment or landmarks. It can be integrated into other tasks and platforms because of the high independence on the specific device. For example, an interesting non-medical application is robotic art and painting restoration performed by Cartesian or collaborative robots. Such robot-based artistic creation also relies on high-precision mechanical motion, which can be supported by AR-based visual feedback and partial scene visualization.

Advanced neural networks have been proposed for different vision tasks with promising performance by evaluating some public or self-made datasets, and it is worth integrating them to implement applications in practice. Hence, we compared the state-of-the-art networks in the stereo reconstruction and segmentation fields to localize the soft tissue in 3D intra-operative space. We adopted the model HSM (Yang et al., 2019) in our case because it shows a reliable depth estimation in terms of both accuracy and inference time. Furthermore, we added a segmentation mask to extract the region of interest from the whole scene. Since the segmentation performance depends on the specific scenes, we captured and annotated an endoscopic dataset based on the dVRK platform in our lab. By referring to the segmentation results of 6-fold cross validation, we found that UNet (Ronneberger et al., 2015) can provide a more reliable segmentation quality compared to other models, so we adopted it in our framework. However, it should be noted that the segmentation quality is influenced by different scenes and training strategies. For example, the transformer-based networks (Chen et al., 2021; Hatamizadeh et al., 2022; Zheng et al., 2021) may provide better estimation with a large number of annotated training images, which need to be compared in the specific tasks. In particular, we evaluated the emerging large model SAM (Kirillov et al., 2023). Although it does not perform better than some other models, it should be noted that this model has a strong generalization ability in different scenes without fine-tuning. With the rise of large models and data, it can be predicted that this kind of model may dominate the vision tasks as it relieves the burden of annotation and training.

By emulating a robot-assisted lymphadenectomy based on the dVRK platform, we obtained some surgical data from ten human subjects. We observed that the proposed AR framework can increase the distance and reduce the number of collisions between the instruments and the delicate blood vessel during operation. Moreover, it does not introduce the extra physical and cognitive load since the overall path and the execution time are similar in two different modalities. Based on the

feedback of the users, they sometimes failed to clamp the lymph nodes due to inaccurate depth perception when operating. Under this circumstance, the color information of the pre-operative model can assist them in judging the proper occasion to clamp the lymph nodes, especially the pink and red colors (the distances are within 1 cm and 0.5 cm, respectively), which can improve their accuracy when clamping the objects. Finally, the average SUS score given by the users increases by 7% when adopting the AR assistance, which means that the proposed AR framework is user-friendly. One limitation comes from the modeling of the surgical instrument. In our experiment, we used the large needle driver to perform the task, and modeling it as a cylinder would introduce small errors in practice. To further improve the modeling accuracy of the instrument, a possible solution is to add an additional neural network to recognize the skeleton of the gripper (Lu et al., 2021), although it will slow down the framework. Another limitation comes from the phantom-based environment for the emulated surgical operation. In clinical practice, the in vivo environment of patients will be more complex and dynamic. The robustness of the neural network-based AR framework remains to be further enhanced (Solak, Faydasicok, & Arik, 2023; Song, Peng, Song, & Stojanovic, 2024; Song, Song, Stojanovic, & Song, 2023).

6. Conclusion

This paper proposes a markerless augmented reality framework to implement pre-operative model visualization on the intra-operative scenes, and it also provides minimum distance detection between instruments and the delicate blood vessel for safety. It can be integrated into other existing robotic platforms and tasks because it does not rely on specific devices. Comprehensive comparison studies are performed to explore the best combination for intra-operative blood vessel reconstruction, and the framework usability evaluation by ten human subjects presents that the proposed framework enhances safety during operation and does not introduce extra burden, which shows its potential in clinical applications.

Augmented reality provides the possibility to enhance surgical safety based on visual feedback, and another popular direction, virtual fixtures, is also possible to enhance safety by providing force feedback. In the next step, we will conduct virtual fixtures based on the dVRK platform and compare the difference between visual feedback and force feedback in surgical assistance.

Compliance with ethical standards

Ethical approval This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Politecnico di Milano Ethics Committee (authorization no. 30/2023).

Informed consent Written informed consent was obtained from all human subjects included in the study.

CRediT authorship contribution statement

Ziyang Chen: Writing – original draft, Validation, Methodology, Conceptualization. **Laura Cruciani:** Methodology, Investigation. **Ke Fan:** Writing – original draft, Conceptualization. **Matteo Fontana:** Investigation, Conceptualization. **Elena Lievore:** Investigation, Conceptualization. **Ottavio De Cobelli:** Supervision, Investigation, Conceptualization. **Gennaro Musi:** Supervision, Investigation, Conceptualization. **Giancarlo Ferrigno:** Supervision, Investigation, Conceptualization. **Elena De Momi:** Supervision, Investigation, Conceptualization.

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.

Data availability

No data was used for the research described in the article.

References

Allan, M., Mcleod, J., Wang, C., Rosenthal, J. C., Hu, Z., Gard, N., et al. (2021). Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133.

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures: vol. 1611*, (pp. 586–606). Spie.

Bianchi, L., Chessa, F., Angiolini, A., Cercenelli, L., Lodi, S., Bortolani, B., et al. (2021). The use of augmented reality to guide the intraoperative frozen section during robot-assisted radical prostatectomy. *European Urology*, *80*(4), 480–488.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Brooke, J., et al. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7.

Chan, J. Y., Holsinger, F. C., Liu, S., Sorger, J. M., Azizian, M., & Tsang, R. K. (2020). Augmented reality for image guidance in transoral robotic surgery. *Journal of Robotic Surgery*, *14*, 579–583.

Chang, J.-R., & Chen, Y.-S. (2018). Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5410–5418).

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, Z., Terlizzi, S., Da Col, T., Marzullo, A., Catellani, M., Ferrigno, G., et al. (2022). Robot-assisted ex vivo neobladder reconstruction: preliminary results of surgical skill evaluation. *International Journal of Computer Assisted Radiology and Surgery*, *17*(12), 2315–2323.

Cheng, X., Zhong, Y., Harandi, M., Dai, Y., Chang, X., Li, H., et al. (2020). Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, *33*.

Cheng, X., Zhong, Y., Harandi, M., Drummond, T., Wang, Z., & Ge, Z. (2022). Deep laparoscopic stereo matching with transformers. In *Medical image computing and computer assisted intervention–mICCAI 2022: 25th international conference, Singapore, September 18–22, 2022, proceedings, part VII* (pp. 464–474). Springer.

Da Col, T., Mariani, A., Deguet, A., Menciassi, A., Kazanzides, P., & De Momi, E. (2020). Scan: System for camera autonomous navigation in robotic-assisted surgery. In *2020 IEEE/RSJ international conference on intelligent robots and systems* (pp. 2996–3002). IEEE.

De Backer, P., Van Praet, C., Simoens, J., Lores, M. P., Creemers, H., Mestdagh, K., et al. (2023). Improving augmented reality through deep learning: Real-time instrument delineation in robotic renal surgery. *European Urology*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Edwards, P. E., Psychogyios, D., Speidel, S., Maier-Hein, L., & Stoyanov, D. (2022). SERV-CT: A disparity dataset from cone-beam CT for validation of endoscopic 3D reconstruction. *Medical Image Analysis*, *76*, Article 102302.

Eigen, D., Puhrsch, C., & Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, *27*.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Garcia, V., Debreuve, E., & Barlaud, M. (2008). Fast k nearest neighbor search using GPU. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 1–6). IEEE.

Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, *33*, 22517–22529.

Geiger, A., Roser, M., & Urtasun, R. (2011). Efficient large-scale stereo matching. In *Computer vision–ACCV 2010: 10th Asian conference on computer vision, queenstown, New zealand, November 8-12, 2010, revised selected papers, part i 10* (pp. 25–38). Springer.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., & Tan, P. (2020). Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2495–2504).

Guo, X., Yang, K., Yang, W., Wang, X., & Li, H. (2019). Group-wise correlation stereo network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3273–3282).

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., et al. (2022). Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 574–584).

Horn, B. K. (1987). Closed-form solution of absolute orientation using unit quaternions. *Journal of the Opticla Society of America A*, *4*(4), 629–642.

Kim, D., & Kim, J. (2023). CT-loc: Cross-domain visual localization with a channel-wise transformer. *Neural Networks*, *158*, 369–383.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026).

Lee, D., Kong, H.-J., Kim, D., Yi, J. W., Chai, Y. J., Lee, K. E., et al. (2018). Preliminary study on application of augmented reality visualization in robotic thyroid surgery. *Annals of Surgical Treatment and Research*, *95*(6), 297–302.

Li, Z., Liu, X., Drenkow, N., Ding, A., Creighton, F. X., Taylor, R. H., et al. (2021). Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 6197–6206).

Li, J., Wang, P., Xiong, P., Cai, T., Yan, Z., Yang, L., et al. (2022). Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16263–16272).

Liu, B., Yu, H., & Qi, G. (2022). Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13012–13021).

Lu, J., Jayakumari, A., Richter, F., Li, Y., & Yip, M. C. (2021). Super deep: A surgical perception framework for robotic tissue manipulation using deep learning for feature extraction. In *2021 IEEE international conference on robotics and automation* (pp. 4783–4789). IEEE.

Moccia, S., De Momi, E., El Hadji, S., & Mattos, L. S. (2018). Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics. *Computer Methods and Programs in Biomedicine*, *158*, 71–91.

Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., et al. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Penza, V., De Momi, E., Enayati, N., Chupin, T., Ortiz, J., & Mattos, L. S. (2017). EnViSoRS: Enhanced vision system for robotic surgery. a user-defined safety volume tracking to minimize the risk of intraoperative bleeding. *Frontiers in Robotics and AI*, *4*, 15.

Qian, L., Wu, J. Y., DiMaio, S. P., Navab, N., & Kazanzides, P. (2019). A review of augmented reality in robotic-assisted surgery. *IEEE Transactions on Medical Robotics and Bionics*, *2*(1), 1–16.

Rodler, S., Kidess, M. A., Westhofen, T., Kowalewski, K.-F., Belenchon, I. R., Taratkin, M., et al. (2023). A systematic review of new imaging technologies for robotic prostatectomy: from molecular imaging to augmented reality. *Journal of Clinical Medicine*, *12*(16), 5425.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234–241). Springer.

Schiavina, R., Bianchi, L., Lodi, S., Cercenelli, L., Chessa, F., Bortolani, B., et al. (2021). Real-time augmented reality three-dimensional guided robotic radical prostatectomy: preliminary experience and evaluation of the impact on surgical planning. *European Urology Focus*, *7*(6), 1260–1267.

Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., et al. (2020). *opencv/cvat: v1. 1.0*. Zenodo.

Shen, Z., Dai, Y., & Rao, Z. (2021). Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13906–13915).

Solak, M., Faydasicok, O., & Arik, S. (2023). A general framework for robust stability analysis of neural networks with discrete time delays. *Neural Networks*, *162*, 186–198.

Song, X., Peng, Z., Song, S., & Stojanovic, V. (2024). Anti-disturbance state estimation for PDT-switched RDNNs utilizing time-sampling and space-splitting measurements. *Communications in Nonlinear Science and Numerical Simulation*, Article 107945.

Song, X., Song, Y., Stojanovic, V., & Song, S. (2023). Improved dynamic event-triggered security control for T–S fuzzy LPV-PDE systems via pointwise measurements and point control. *International Journal of Fuzzy Systems*, *25*(8), 3177–3192.

Song, X., Wu, N., Song, S., Zhang, Y., & Stojanovic, V. (2023). Bipartite synchronization for cooperative-competitive neural networks with reaction–diffusion terms via dual event-triggered mechanism. *Neurocomputing*, *550*, Article 126498.

Su, H., Hu, Y., Karimi, H. R., Knoll, A., Ferrigno, G., & De Momi, E. (2020). Improved recurrent neural network-based manipulator control with remote center of motion constraints: Experimental results. *Neural Networks*, *131*, 291–299.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Venugopal, A., Moccia, S., Foti, S., Routray, A., MacLachlan, R. A., Perin, A., et al. (2022). Real-time vessel segmentation and reconstruction for virtual fixtures for an active handheld microneurosurgical instrument. *International Journal of Computer Assisted Radiology and Surgery*, *17*(6), 1069–1077.

Wendler, T., van Leeuwen, F. W., Navab, N., & van Oosterom, M. N. (2021). How molecular imaging will enable robotic precision surgery: The role of artificial intelligence, augmented reality, and navigation. *European Journal of Nuclear Medicine and Molecular Imaging*, *48*(13), 4201–4224.

Williams, F. (2022). Point cloud utils. https://www.github.com/fwilliams/point-cloud-utils.

Xu, G., Cheng, J., Guo, P., & Yang, X. (2022). Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12981–12990).

Xu, G., Wang, X., Ding, X., & Yang, X. (2023). Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Yang, G., Manela, J., Happold, M., & Ramanan, D. (2019). Hierarchical deep stereo matching on high-resolution images. In *The IEEE conference on computer vision and pattern recognition*.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(11), 1330–1334.

Zhao, C., Sun, Q., Zhang, C., Tang, Y., & Qian, F. (2020). Monocular depth estimation based on deep learning: An overview. *Science China Technological Sciences*, *63*(9), 1612–1627.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., et al. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6881–6890).

Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2019). Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, *39*(6), 1856–1867.

Zhuang, J. (2018). LadderNet: Multi-path networks based on U-net for medical image segmentation. arXiv preprint arXiv:1810.07810.