

A Preliminary Exploration to Make Stereotactic Surgery Robots Aware of the Semantic 2D/3D Working Scene

Liang Li¹, Pengfei Feng, Hui Ding, *Member, IEEE*, and Guangzhi Wang², *Member, IEEE*

Abstract—Scene perceptual ability is key to developing autonomy and intelligence in surgical robots. This study helps stereotactic surgical robots detect and segment key objects in unstructured surgical scenes. First, we construct a neurosurgery robot working scene dataset. Next, we propose a 2Dimage-scene-aware pipeline that integrates a Mask R-CNN (mask region-based convolutional neural network) with a conditional random field and a superpixel method; the pipeline detects and segments key objects, such as the patient's head, head frame, and body. Then, we establish a multiview projection voting and supervoxel fusion pipeline that extracts further information from a 3D point cloud scene. The proposed method was tested in different clinical scenarios, and the results show that the method can detect and segment specific surgical objects and achieves comparable accuracy and stability on both 2D images and 3D point cloud data. The average precision (AP) and average 2D and 3D Dice scores for the patient's head were 97.65, 91.6, and 92.6, respectively. Better segmentation performances can be achieved when the data-based neural network method further integrates the traditional color and contour-based image processing methods. The proposed solution allows stereotactic surgical robots to better understand their surroundings, provides semantic information useful for subsequent tasks, and lays a foundation for autonomous stereotactic surgical robots.

Index Terms—Stereotactic surgical robot, surgical scene awareness, robot vision, semantic segmentation, information fusion.

I. INTRODUCTION

ROBOTS have always been imagined as synonymous with high autonomy and high intelligence. However, most robots in the surgical field play a role only in an instrument with enhanced dexterity; their autonomy and intelligence are still poor [1], [2]. One main challenge for autonomous surgical robots is to give them the ability to perceive unstructured

surgical scenes. Therefore, an increasing number of studies have focused on improving the scene intelligence of surgical robots [3]. However, most of these studies have been applied to robots used in endoscopic surgery, such as for instrument tracking [4], endoscopic scene reconstruction [5], and endoscopic surgical process recognition [6]. We argue that for stereotactic surgical robots [7]–[9], it is also necessary to study the problem of surgical scene perception and understanding.

Stereotactic surgical robots face the problem of an unstructured working environment. Patients have unique anatomical structures and surgical plans and can appear in a variety of postures, in head frames, with ventilator tubes, and with other instrument connections in various locations in the robot workspace. For the robot to work, the surgical plan needs to be registered to the physical environment, and the robot arm movements must avoid collisions with the above surrounding objects. However, because robots cannot perceive and understand these unstructured working scenes, current stereotactic robotic surgery still relies on human operators to plan and manipulate the robot's movements to meet the above requirements. Moreover, due to human factors such as fatigue, inattention, and inexperience, such manual operations not only make surgical procedures more cumbersome and time consuming but also affect surgery consistency and quality and increase the risks of surgical procedures.

Research on providing stereotactic surgery robots with the ability to understand the working scene is limited. In the ROBOCAST project in 2010, Ferrigno *et al.* integrated multiple sensing information into keyhole neurosurgery robots to realize automatic path planning. This project used optical markers attached to key surgical objects to help make robots aware of the objects' positions in the operating room [10]. A similar study used radio-frequency identification (RFID) tags to assist the computer in understanding the operating room environment [11]. Although the presence of such artificial markers allows robots to achieve object detection and position tracking, these markers do not provide object shape information. Moreover, it is unrealistic to manually mark all the objects in the environment. In 2016, Beyl *et al.* used both time-of-flight (TOF) and Kinect cameras to acquire point clouds and achieve personnel detection in the operating room [12]. The study used no artificial markers; however, the method used in the study can identify only a single person and does not provide a means of understanding general scene object information. Thus, making stereotactic surgical

Manuscript received July 20, 2020; revised January 6, 2021 and September 2, 2021; accepted October 15, 2021. Date of publication October 29, 2021; date of current version February 22, 2022. This article was recommended for publication by Associate Editor X. Luo and Editor P. Dario upon evaluation of the reviewers' comments. This work was supported in part by the Ministry of Science and Technology of China under Grant 2019YFC0119503 and Grant 2017YFA0205904; in part by the Tsinghua University Initiative Scientific Research Program under Grant 20197010009; and in part by the Beijing Science and Technology Research Program under Grant Z191100007619036. (Liang Li and Pengfei Feng are contributed equally to this work.) (Corresponding author: Guangzhi Wang.)

The authors are with the Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing 100084, China (e-mail: lil17@mails.tsinghua.edu.cn; fpf19@mails.tsinghua.edu.cn; dinghui@tsinghua.edu.cn; wgz-dea@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TMRB.2021.3124160

2576-3202 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See <https://www.ieee.org/publications/rights/index.html> for more information.

robots visually understand natural objects, such as surgeons in surgical scenes, remains unsolved.

Deep learning technology has shown enormous potential in scene-understanding tasks [13]. Many scene-understanding applications have been developed; these include automatic traffic scenes [14], indoor scenes [15], and endoscopic scenes [4]. Recently, the application potential of deep learning-based methods in medical scenarios, such as intensive care units (ICUs) and operating rooms, has also attracted increasing attention from researchers [16], [17]. However, in general, the application of computer vision and deep learning technology in the operating room still focuses on traditional surgical requirements, such as procedure duration prediction, gesture recognition, workflow recognition and tool tracking in endoscopic scenes [2], [18], [19]. Understanding surgical scenes is a fundamental aspect of autonomous surgical robotic systems and is the first step in developing advanced control or path planning techniques for medical robots. There is very limited research on using deep learning to help robots understand scene information in operating rooms. Most studies have focused on the endoscopic surgical scene [4]–[6]. Only one study attempted to use deep learning to enable Da Vinci surgical robots to understand operating room scenes [20].

From a technical point of view, semantic segmentation is the basic task of scene understanding. There are many studies applying deep learning techniques to perform 2D image scene semantic segmentation [21]. Although many newer methods have been proposed, in terms of recognition and segmentation accuracy, the mask region-based convolutional neural network (Mask R-CNN) [22], proposed by He *et al.* in 2017, still maintains a leading edge even in 2020 [23].

Despite the emergence of deep learning methods, 3D point cloud scene understanding remains a challenging task due to problems such as occlusion, viewpoint variations, scale changes, and disorderly arrangement [24]. The ongoing unstructured point cloud-based learning method [25] has shown great potential, but a recognized and stable learning network has not been produced [26], especially for personalized surgical robot point cloud data in the operating room. Another type of voxel-grid-based 3D CNN point cloud understanding method [27] has been reported but is computationally expensive [24]. The multiview projection-based method [28] uses the geometric relationship between the 3D point cloud and the 2D image to map the semantic information from a 2D image to a 3D point cloud. This method can take advantage of mature high-performance 2D image processing networks but will inevitably lose some information when projecting the 3D point cloud to the 2D image [26]. Moreover, due to the uninterpretable characteristics, there are still potential risks that cannot be ignored in the medical application of methods based entirely on learning [29]. For medical applications, learning-based methods need to be further explored in a safer and interpretable direction.

The availability of effective training data is both a prerequisite and the basis for these methods. Currently, however, no related dataset exists for stereotactic surgical robot working scenes; therefore, it is difficult to explore applications of these data-based methods for stereotactic surgical robot scene

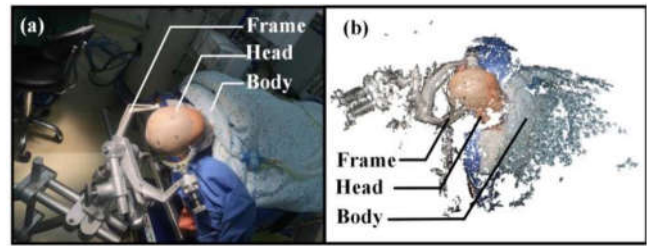


Fig. 1. Examples of neurosurgical robot surgery scene data: (a) 2D surgical scene image; (b) 3D surgical scene point cloud corresponding to the 2D image.

understanding. Moreover, the current methods are mostly aimed at purely network-based scene understanding, and stable 3D-scene-understanding methods are not yet mature [17]. Thus, achieving scene understanding for stereotactic surgical robots requires further research, and the target objects and requirements for stereotactic robot scene understanding also need to be further defined.

The main goal of this paper is to find a method by which stereotactic surgery robots can understand unstructured surgery scenes from 2D and/or 3D visual data. We adopt the understanding of a stereotactic neurosurgical robot scene as an example, combine deep learning and traditional visual information processing methods, construct a stereotactic neurosurgery robot working environment dataset, and finally detect and segment typical surgical objects of interest (such as the patient's head, head frame, and body) in surgical scenes from 2D images and 3D cloud point data. We hope this work will lay a foundation for subsequent stereotactic surgery robot scene intelligence studies.

II. METHOD

A. Overview

1) *Visual Scene Data*: In previous works, we constructed several stereotactic surgical robot systems with flexible hand-eye configurations. These systems obtain either 2D scene images from the perspective of the end or joint of the robot arm [30], [31] or can acquire 3D point clouds of surgical scenes through multiview reconstruction [31] or structured light [32] algorithms. To ensure research continuity, we applied these 2D and 3D scene data to study the surgical-scene-understanding problem. Fig. 1 shows some example data, including both a 2D image and a visualized 3D point cloud scene.

2) *Scene-Understanding Objects*: During surgery, stereotactic surgical robots are usually placed at the top side of the patient's head and facing the patient. A head frame is used to affix the patient's head to the robot in a suitable pose. Typically, the patient's head, head frame, and body are the main objects in the neurosurgical robot's working scene. As a preliminary exploration of stereotactic surgery robot scene understanding and based on our understanding of neurosurgery robots' clinical needs, in this study, the patient's head, head frame, and body were chosen as the objects that the robot needs to understand. To meet the demands for intelligent robot tracking, registration, and path planning, we want the method

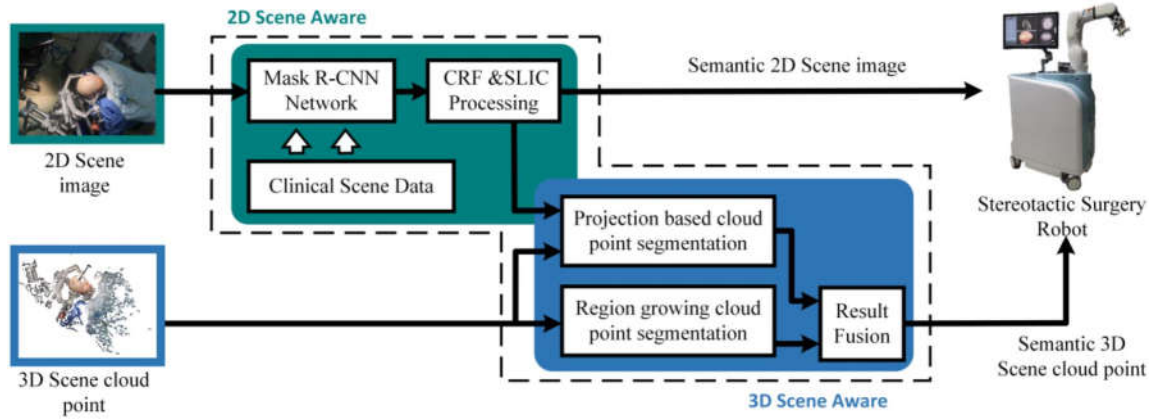


Fig. 2. Scene understanding framework for stereotactic surgical robots.

to detect and segment these objects and obtain their semantic information.

3) *Framework of the Scene-Understanding Process*: Unstructured factors, such as personalized patient anatomy, various intraoperative poses, lighting conditions, and object occlusions, are the main challenges for neurosurgery robot scene understanding. The framework shown in Fig. 2 was designed to address this problem. This framework contains two main parts: 2D scene-aware and 3D scene-aware components. The 2D scene-aware components include a Mask R-CNN network module that aims at learning experience from the clinical scene datasets to detect and roughly segment the target objects. Then, a conditional random field (CRF) and simple linear iterative clustering (SLIC) processing module is used to finely segment the target objects. The 3D scene-aware components include a projection-based point cloud segmentation module that uses the 2D scene-aware results to realize preliminary recognition and segmentation for 3D point cloud scenes. Additionally, a regional growth point cloud segmentation module is used to introduce the point cloud's color information. Then, the final finely segmented target object point cloud is obtained by fusing the above two segmentation results. After the 2D scene-aware and 3D scene-aware processes have finished, their respective results are input into the stereotactic robot system for subsequent tasks, such as registration and path planning.

B. 2D Scene Understanding

1) *Local Data Collection*: We built a local 2D surgical scene dataset for neural network training. Surgical-scene photos were collected by the SINO neurosurgery robot (Sinovation Medical Technology Co., Ltd., Beijing, China) in the operating rooms of different hospitals. The main view angles of the selected photos are from the end and joint areas of the robot arm.

2) *Mask R-CNN-Based Object Detection and Fusion-Based Segmentation*: The 2D scene understanding framework is shown in Fig. 3 and includes two stages: network prediction and information fusion. In the network prediction stage, we train a Mask R-CNN network [22] to identify the region of

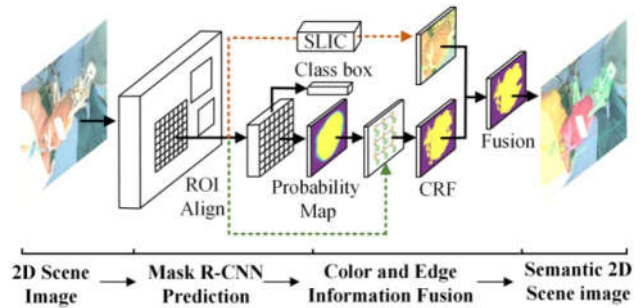


Fig. 3. 2D scene understanding framework.

interest (ROI) and provide a category label and a segmentation probability map for each targeted object. The second stage uses the conditional random field (CRF) for fine segmentation based on the segmentation probability map obtained from the Mask R-CNN network. Finally, the CRF segmentation results are fused using the simple linear iterative clustering (SLIC) superpixel segmentation result to obtain precise segmentation boundaries.

Both pretraining and a data augmentation strategy were used to prevent the Mask R-CNN model from underfitting or overfitting. The network was pretrained on the Common Objects in Context (COCO) dataset [33] and then fine-tuned on our local dataset. Using this approach, the network learns general visual features from the large-scale public visual dataset and then learns the specific features only from the local dataset, thereby preventing underfitting problems. To avoid overfitting, we enhanced the local dataset by using online data augmentation. In each training epoch, we randomly select a certain proportion of the image data and perform random cropping, flipping, affine transformations, blurring, contrast changes, brightness adjustments, grayscale adjustments, and other augmentation operations to simulate scene data under as many conditions as possible.

The CRF fine segmentation uses fully connected CRFs with Gaussian edge potentials [34]. The Gibbs energy of the CRF is calculated by Equation (1), where unary potential $\psi_u(x_i)$ is used to give the prior segmentation result, set to

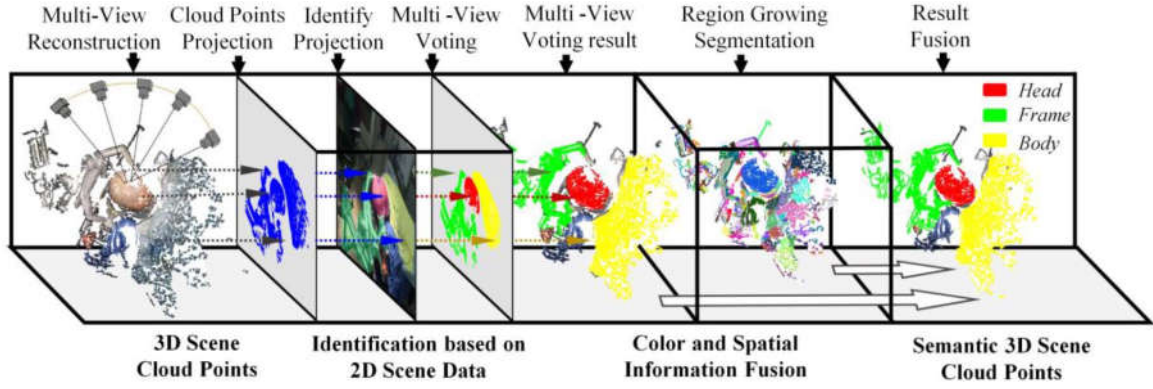


Fig. 4. 3D point cloud scene understanding framework based on multiview projection voting and supervoxel segmentation fusion.

$\psi_u(x_i) = -\log P(x_i)$, where $P(x_i)$ represents the probability that a pixel belongs to a certain category, provided by the FCN module of the Mask R-CNN. The pairwise potentials $\psi_p(x_i, x_j)$ are used to introduce pixel color and positional information to assist with the fine segmentation. As shown in Equation (2), there are two contrast-sensitive kernel functions for pixel color vector I_i and position vector p_i . The first function (an appearance kernel) tends to classify nearby pixels with similar colors into the same class, while the second function (a smoothness kernel) is used to penalize smaller isolated regions. Here, θ_α , θ_β and θ_γ are the control parameters for the corresponding items; $\omega^{(1)}$ and $\omega^{(2)}$ are linear Gaussian combination weights; and $\mu(x_i, x_j)$ is the label compatibility function given by the Potts model. Specifically, $\mu(x_i, x_j) = 1$ when $x_i \neq x_j$; otherwise, $\mu(x_i, x_j) = 0$:

$$E(X) = \sum_i \psi_u(x_i) + \sum_{i < j} \psi_p(x_i, x_j) \quad (1)$$

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \left[\underbrace{\omega^{(1)} \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + \underbrace{\omega^{(2)} \exp\left(-\frac{\|p_i - p_j\|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}} \right]. \quad (2)$$

The superpixels are generated by the SLIC algorithm [35]. Algorithm 1 shows the fusion strategy for the superpixels and the CRF segmentation results. The algorithm counts the pixel labels in each superpixel block on the premise of exceeding the minimum acceptance overlap ratio threshold; then, the algorithm selects the label with the largest overlap ratio as the label for the entire superpixel block.

C. 3D Scene Understanding

The 3D-scene-understanding framework is shown in Fig. 4. First, a multiview projection voting method is constructed to detect objects in the point cloud according to the 2D scene

Algorithm 1 Superpixel Segmentation Fusion

Input:

Segmented superpixel $\{S_p\}_{k=1,2,3,\dots,K}$
 Segmented CRF mask $\{Mask^c\}_{N,c=1,2,3,\dots,Nc}$
 K: Total number of superpixels
 Nc: Category code of each object

Input parameter:

Acceptance overlap ratio — $overlap_{accept}$

Output:

Mask after fusion $\{FusionMask_k^c\}_{k=1,2,3,\dots,N,c=1,2,3,\dots,Nc}$

Step1-Calculate overlap ratio

For each superpixel: $k = 1, 2, 3, \dots, K$
 Count overlap ratio of each pixel category:
 $ratio_k^c = \text{length}(S_p \cap Mask^c) / \text{length}(S_p)$
 $c = 1, 2, 3, \dots, Nc$

Step2-superpixel category determination

For each superpixel: $k = 1, 2, 3, \dots, K$
 If $\max(\{ratio_k^c\}_{c=1,2,3,\dots,Nc}) > Overlap_{accept}$
 $FusionMask_k^c = \arg \max(\{ratio_k^c\}_{c=1,2,3,\dots,Nc})$

Return Fusion mask $\{FusionMask_k^c\}_{k=1,2,3,\dots,N,c=1,2,3,\dots,Nc}$

image understanding result. Then, the method further considers the color information in the point cloud. The region growing-based supervoxel segmentation result is fused with the first-stage multiview projection result, and finally, fine 3D point cloud object semantic segmentation is achieved.

1) *Point Cloud Projection and Semantic Recognition*: The mapping relationship between the 3D scene point cloud and the 2D scene image pixels is established through a camera projection matrix. As shown in Equation (3), where K is the camera's intrinsic matrix, RT^p is the camera's extrinsic matrix at viewing angle p , and X_i , Y_i and Z_i are the spatial coordinates of the i -th cloud point. Here, u_i^p and v_i^p represent the pixel position of a point projected to the 2D scene image under the viewing angle p . The pixel category label is then determined based on the location (u_i^p, v_i^p) through the neural network trained as described above. Thus, a point cloud label corresponding to the pixel position can be inversely deduced.

$$z \begin{bmatrix} u_i^p \\ v_i^p \\ 1 \end{bmatrix} = K \bullet RT^p \bullet \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}. \quad (3)$$

Algorithm 2 Multi-view Cloud Point Recognition Voting With Personalized Voting Weight

Input:

3D scene cloud point : $\{P_i\}_{i=1,2,3,\dots,N}$
 Annotated 2D scene image: $\{I_C^p\}_{p=1,2,3,\dots,Np, C=1,2,3,\dots,Nc}$
 Projection matrix of 2D scene image: $\{K \cdot RT^p\}_{p=1,2,3,\dots,Np}$
 N: Total number of point clouds
 Np: Total number of projection views
 Nc: Category code of each object

Input parameter:

Personalized weights— $\{weight_C\}_{C=1,2,3,\dots,Nc}$
 Acceptance threshold— $Threshold_{accept}$

Output:

Category index of 3D scene point cloud: $\{PointIndex_i\}_{i=1,2,3,\dots,N}$

Step1-Initialization

Initialize vote count variable: $\{Vote_i^C = 0\}_{i=1,2,3,\dots,N, C=1,2,3,\dots,Nc}$
 Initialize point category index: $\{PointIndex_i = -1\}_{i=1,2,3,\dots,N}$

Step2-Calculate cloud points 2D projection position

For each angle of view: $p = 1, 2, 3, \dots, Np$
 $projection_i^p = Round(K \cdot RT^p \cdot P_i), i = 1, 2, 3, \dots, N$

Step3-Multi-view voting

For each angle of view: $p = 1, 2, 3, \dots, Np$
 For each category: $C = 1, 2, 3, \dots, Nc$
 If project position belong to the category: $projection_i^p \in I_O^p$
 Votes plus: $vote_i^C = Vote_i^C + 1, i = 1, 2, 3, \dots, N$

Step4-Voting correction with personalized weights

For each category: $C = 1, 2, 3, \dots, Nc$
 $vote_i^C = Vote_i^C * Weight_C, i = 1, 2, 3, \dots, N$

Step5-Point cloud category determination

For each point: $i = 1, 2, 3, \dots, N$
 If $\max(\{Vote_i^C\}_{C=1,2,3,\dots,Nc}) > Threshold_{accept}$
 $PointIndex_k = \arg \max(\{Vote_i^C\}_{C=1,2,3,\dots,Nc})$

Return scene point cloud category index $\{PointIndex_i\}_{i=1,2,3,\dots,N}$

2) *Multiview Projection Voting*: A view from a single projection angle cannot separate overlapping point clouds. Classification from multiple angles of view could solve this problem, but imperfect segmentation in the object boundary region may occur in 2D scene images, thus resulting in further conflicting classification results at the object's 3D point cloud boundary. We introduce a multiview voting mechanism to overcome this problem, as shown in Algorithm 2. The algorithm projects each point in the 3D scene into all the original reconstructed 2D scene images and then votes on the object attributions of points according to the 2D scene segmentation result. The number of votes that a point receives for each object type is counted separately. The final point category is determined by the highest number of votes that exceeds a predetermined acceptance threshold. The acceptance threshold $Threshold_{accept}$ is proportional to the number of view angles participating in the voting, as shown in Equation (4), where Num_{min} is the minimum acceptable threshold, Num_{all} is the total number of voting perspectives, and λ is a correlation coefficient between the total number of viewing angles; λ has a range from 0 ~ 1:

$$Threshold_{accept} = Num_{min} + \lambda \cdot (Num_{all} - Num_{min}). \quad (4)$$

3) *Personalized Projection Voting Weight*: Neural networks achieve different segmentation and detection performance levels for objects of different complexity, thereby resulting in

unbalanced segmentation accuracy for different objects during the 3D point cloud projection segmentation voting. The strategy of personalizing weights is designed to alleviate this problem. An independent weight coefficient is assigned to each object. This weight coefficient is used to weight the projected voting result for each object. Then, the result is counted as the final voting value for point recognition. A lower weight coefficient clearly results in object point cloud segmentation with higher accuracy but lower integrity. With low weight coefficients, higher projection votes are necessary to reach above the acceptance threshold. In contrast, a high weight coefficient means that an object can reach the acceptance classification threshold with fewer votes. Although the risk of misclassification may increase, more point clouds belonging to the category are obtained. With such a personalized weighting mechanism, we can design personalized weights for different objects to obtain a desired 3D point cloud segmentation result.

Here, we provide a method to automatically calculate the object's personalized weight value according to the network's segmentation and recognition performance in the 2D scene. As shown in Equation (5), the network's average precision (AP) for detecting the target object is used to compensate for the differences in detection performance, while for segmentation accuracy, we use the intersection over union (IOU) index to compensate for the differences in segmentation performance. According to Equation (5), the lower the 2D segmentation accuracy and detection precision are, the higher the weight of the votes is; thus, the final number of votes increases accordingly. In this way, for objects with low 2D-understanding performance, the point cloud object type can be determined by fewer projection votes; this helps ensure the integrity of the object classification as much as possible and ensures the fairness of the point cloud attribution in critical areas that compete with other objects with high 2D segmentation accuracy.

$$\begin{aligned} &Weight_{Head} : Weight_{Frame} : Weight_{Body} \\ &= 1 : \frac{AP_{Head}^{50} \cdot IOU_{Head}}{AP_{Frame}^{50} \cdot IOU_{Frame}} : \frac{AP_{Head}^{50} \cdot IOU_{Head}}{AP_{Body}^{50} \cdot IOU_{Body}}. \end{aligned} \quad (5)$$


4) *Fusion of Supervoxel Segmentation*: The underlying idea of the supervoxel segmentation result fusion method is generalized from the superpixel segmentation fusion strategy in 2D-understanding tasks (shown in Algorithm 1). The color-based 3D point cloud region growing algorithm was first used to segment the scene point cloud into supervoxels [36]. Then, the cloud points were further assigned to different categories by the above multiview projection voting method. The percentage of each point category in the supervoxel is calculated as the overlap ratio. For each supervoxel, the category with the highest overlap ratio was selected as the final semantic label of the entire supervoxel block.

III. RESULTS

This section first presents a description of the dataset and then reports the overall results of the proposed 2D-/3D-scene-understanding method. Finally, we show some additional results from various method generalizability aspects.

TABLE I
DATA SET PARTITION AND TYPICAL EXAMPLES

Data set	Hospitals ID	Number of scenes	Number of images
Training set	#1~#6	35	1019
Validation set	#1~#6	8	232
Test set 1	#1~#6	8	209
Test set 2	#7~#8	4	118



A. Training and Testing Datasets

The data used in this study were provided by Sinovation Medical Technology Co., Ltd. (Beijing, China). The dataset contains 55 robot-assisted deep brain stimulation (DBS) or stereoelectroencephalography (SEEG) surgical scenes from 8 hospitals; these scenes comprise 1,578 photos. All the obtained data were approved by the relevant hospital ethics committees.

All image data were collected during the robot registration preparation stage. The sampling timing is manually controlled, and a certain scale of translation and rotation are performed in the camera before each photo is collected. In the clinic, a robot may encounter certain factors, such as an obscured robot target, an object that exceeds the robot's field of view, and lighting changes in the operating room; therefore, the randomness of the photo viewing angle, shooting distance, and lighting conditions were ensured as much as possible during data collection.

The data from 8 hospitals were divided into two categories, as shown in Table I. The data from 6 hospitals were used to construct a training set, a validation set and test set 1. The data from the remaining two hospitals were used to construct test set 2. The training set contains 35 surgical scenes comprising 1,019 photos. The validation set contains 8 surgical scenes comprising 232 photos. Test set 1 contains 8 surgical scenes comprising 209 photos. The training set, validation set, and test set 1 are randomly divided by surgical scene unit such that each of these datasets is guaranteed to contain surgical scenes from all 6 hospitals. Test set 2 contains 4 scenes comprising 118 photos and is used to test the generalizability of the proposed method. Moreover, all the photos in test sets 1 and 2 were reconstructed by the structure from motion (SfM) method [37], [38] and the multiview dense reconstruction algorithm [39] to obtain scene point clouds for testing the 3D-scene-understanding method.

B. Results of 2D-Scene Understandings

We trained and tested the network on the corresponding datasets by using a Tesla P100 graphics processing unit (GPU). The backbone of the Mask R-CNN model is a residual network-101 (ResNet101) network. All the training and test images were unified to a size of 1024×1024 . The training

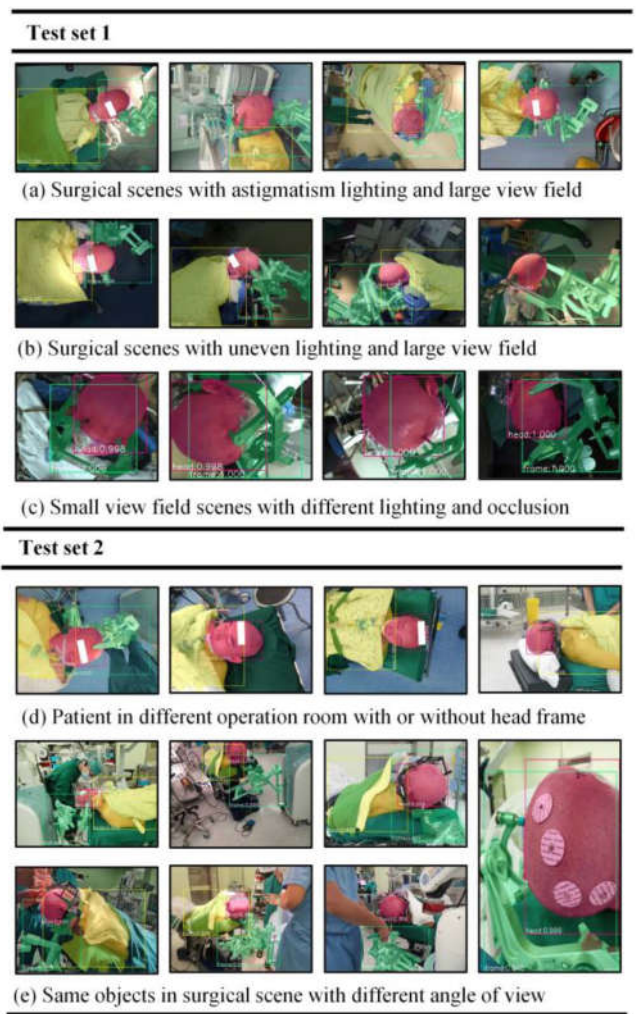


Fig. 5. Results of 2D scene understanding in different unstructured factors selected from test set 1 and test set 2.

is halted just before the loss function curve of the verification set begins to rise. The parameters for the SLIC and CRF algorithms in the fine segmentation stage are manually adjusted in advance.

Fig. 5 intuitively shows the results of 2D-scene understanding on test sets 1 and 2. The method can detect and segment these target objects in 2D scenes under different patient poses, surgical scenes, lighting conditions, and viewing angles. Tables II and III list the quantitative detection and segmentation scores on test sets 1 and 2; in these tables, “Mask R-CNN” indicates that only the Mask R-CNN network was used, while “Fusion” indicates that the Mask R-CNN network, CRF, and SLIC were used. After calculating the statistics, the method obtains average scores of 83.0, 68.6, and 80.8 for AP, IOU, and Dice, respectively, on these two datasets. For the most important robotic surgery object (the patient's head), the method achieves an AP score above 95.8 and a segmented Dice score above 89.4.

To verify the method's ability to segment edge details, the boundary F-scores at the 3- and 5-pixel levels were further calculated following the method in [40]. The results are listed in

TABLE II
QUANTITATIVE SCORES OF 2D SCENES UNDERSTANDING IN TEST SET 1

Objects	Method	Detect ^a AP ⁵⁰	Mask IOU	Mask Dice	^b BF ^{3pixel} scores	BF ^{5pixel} scores
Head	Mask R-CNN	99.5	88.4	93.8	36.0	53.2
	Fusion	99.5	88.2	93.7	50.2	61.1
Frame	Mask R-CNN	90.1	63.5	77.7	15.1	23.7
	Fusion	90.1	63.0	77.3	18.1	27.5
Body	Mask R-CNN	82.2	71.1	83.1	11.5	19.1
	Fusion	82.2	68.1	81.0	12.8	21.1

^aAP50 is the evaluation index of detection precision, when the prediction frame and the label frame overlap by more than 50%, the target is considered to be detected. ^bBF^{3pixel} scores means the boundary F-scores under 3 pixels, BF^{5pixel} scores means the boundary F-scores under 5 pixels

TABLE III
QUANTITATIVE SCORES OF 2D SCENES UNDERSTANDING IN TEST SET 2

Objects	Method	Detect ^a AP ⁵⁰	Mask IOU	Mask Dice	^b BF ^{3pixel} scores	BF ^{5pixel} scores
Head	Mask R-CNN	95.8	81.2	89.6	33.1	47.1
	Fusion	95.8	80.8	89.4	48.7	58.3
Frame	Mask R-CNN	57.8	55.9	71.7	13.8	21.4
	Fusion	57.8	55.1	71.1	25.7	33.3
Body	Mask R-CNN	72.4	61.4	76.1	11.4	17.8
	Fusion	72.4	56.6	72.3	9.7	15.0

^aAP50 is the evaluation index of detection precision, when the prediction frame and the label frame overlap by more than 50%, the target is considered to be detected. ^bBF^{3pixel} scores means the boundary F-scores under 3 pixels, BF^{5pixel} scores means the boundary F-scores under 5 pixels

Tables II and III, which show that the fusion method improves the head object segmentation boundary score by an average of 28.9% and the boundary score of the head frame object segmentation by an average of 41.3%. Fig. 6 compares the results of 2D scene segmentation by different methods; the comparison intuitively reveals the superiority of the fusion method's segmentation performance.

We also measured the algorithm's time efficiency. For an image with an input size of 1024×1024 , the Mask R-CNN network detection takes approximately 0.3 s on average, while CRF and SLIC processing add approximately 0.6 s on average.

C. Results of 3D Scene Understanding

The 3D-scene-understanding performance was tested on the scene point clouds generated by test sets 1 and 2. The projection matrix from the point clouds to each 2D image was estimated by the SFM method [37], [38]. Fig. 7 shows the intuitive results on test sets 1 and 2. The proposed method can effectively detect and segment target objects for different patients, different scenes, different parts, and different qualities of 3D scene point clouds. The quantitative evaluation was conducted using precision, recall, 3D IOU, and 3D Dice metrics. The segmentation gold standard was generated by manually annotated 2D images and the multiangle projection method mentioned above. Tables IV and V show the quantitative scores for each object for test sets 1 and 2, respectively; in these

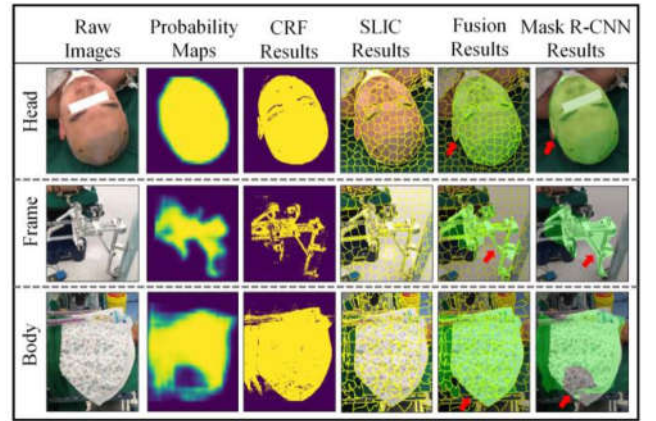


Fig. 6. Result comparison of the fusion and Mask R-CNN segmentation methods. The first column shows the raw image; the second column shows the segmentation probability map output by the Mask R-CNN network; the third column is the result of CRF processing based on the segmentation probability map; the fourth column is the superpixel segmentation result on the raw image using the SLIC algorithm; the fifth column is the mask after fusing of the CRF postprocessing and SLIC; and the sixth column is the mask obtained from the Mask R-CNN probability map using the 50% threshold. The red arrow points to the typical optimization area.

TABLE IV
3D SCENE UNDERSTANDING PERFORMANCE ON TEST SET 1

Objects	Method	Precision	Recall	3D IOU	Dice
Head	Projection	98.6	73.0	72.0	82.8
	Fusion	97.8	91.7	89.8	94.5
Frame	Projection	83.8	60.8	51.0	60.5
	Fusion	80.6	72.1	59.1	65.3
Body	Projection	97.5	63.6	62.0	67.7
	Fusion	95.0	68.9	65.4	69.5

TABLE V
3D SCENE UNDERSTANDING PERFORMANCE ON TEST SET 2

Objects	Method	Precision	Recall	3D IOU	3D Dice
Head	Projection	96.3	79.9	77.1	86.9
	Fusion	93.8	89.6	83.0	90.6
Frame	Projection	97.9	45.8	45.2	61.4
	Fusion	97.5	65.7	64.7	78.6
Body	Projection	96.5	61.9	60.5	73.7
	Fusion	94.0	83.6	79.3	88.2

tables, "Projection" indicates that only the multiview projection method was used, and "Fusion" indicates the result of the region-growing fusion method based on the multiview projection result. The fusion method achieves an average IOU of 73.6 and an average Dice score of 81.1 on all the target objects on test sets 1 and 2. Compared with the projection-only method, the fusion method increases the IOU by 20% and the Dice score by 12% on average. In addition, the tables show that the point cloud of the patient's head in the 3D method achieves good segmentation accuracy, similar to that of the 2D-scene-understanding method.

D. Generalization Performance Analysis

Test set 2 comes from the other two hospitals, data from which are not included in the training and verification datasets.

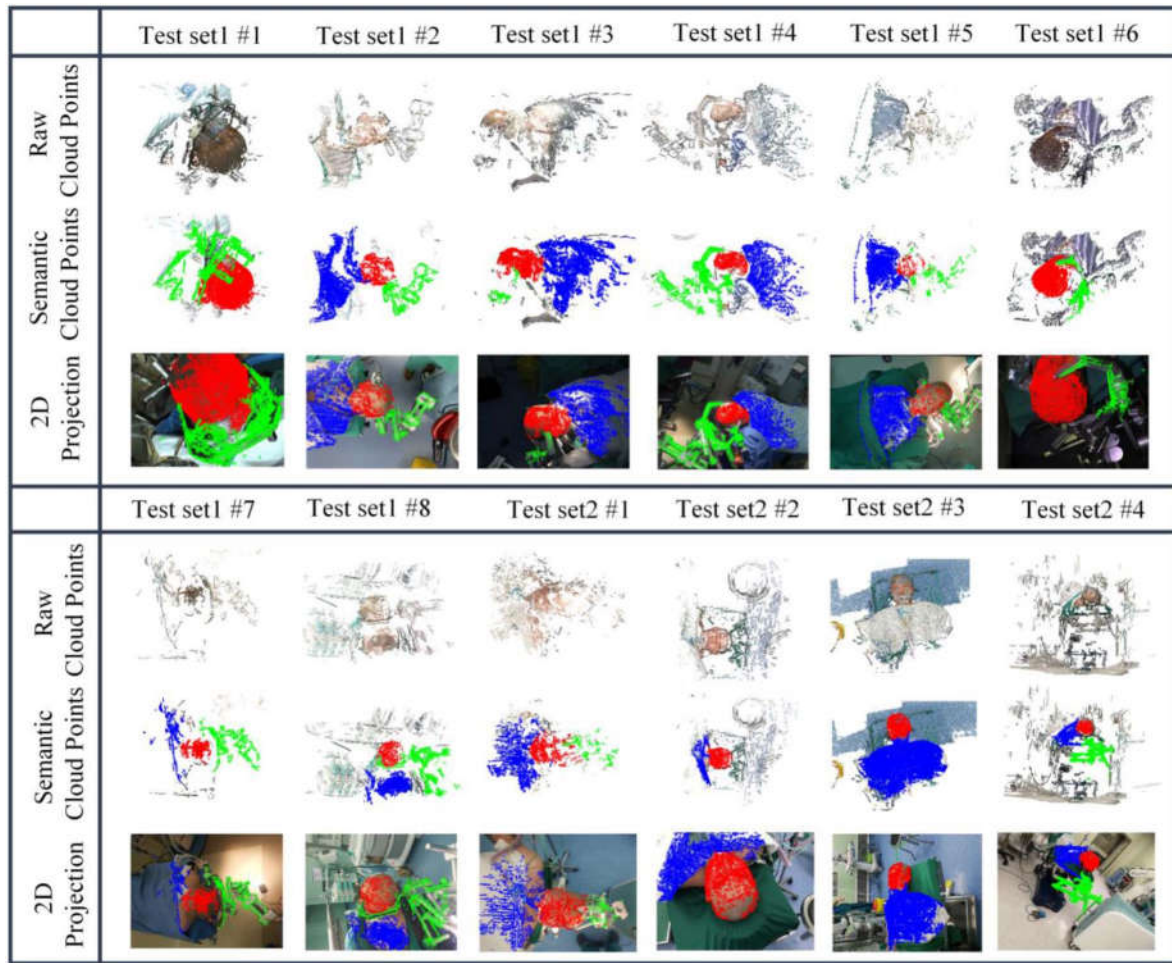


Fig. 7. The results of 3D point cloud scene understanding, including all #1~8 scenes in test set 1 and all #1~4 scenes in test set 2. The first row shows the original point cloud, the second row shows is the semantically segmented point cloud, and the third row shows the scene point cloud projected at the angle of the 2D image view.

Thus, for the network, test set 2 is an unfamiliar working scene that the network has never seen before. Tables II–V show a performance comparison of the 2D- and 3D-scene-understanding methods on test sets 1 and 2, respectively. Fig. 8 provides a more intuitive comparison, where the detection average precision (Fig. 8a), 2D segmentation IOU (Fig. 8b), and 3D segmentation IOU (Fig. 8c) of each scene on test sets 1 and 2 are shown. It can be concluded that 1) the evaluation score of the target objects indeed declines to a certain degree on test set 2 but only slightly. 2) The proposed algorithm achieves high accuracy and robustness in patient head detection and segmentation, and the generalizability of the algorithm in understanding the patient's head in different scenes is the best. 3) However, the detection and segmentation performances with respect to the head frame and the patient's body are relatively low, and the generalization performance varies by scene.

E. Effectiveness of Personalized Voting Weights

To verify the validity of Equation (5), the 3D IOU scores of each object point cloud segmentation under different voting

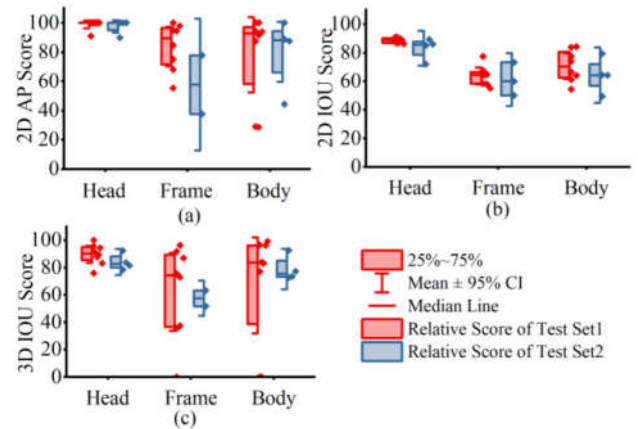


Fig. 8. Generalizability tests on test set 1 and test set 2: (a) comparison of 2D detection AP; (b) comparison of 2D segmentation Dice scores; (c) comparison of 3D segmentation Dice scores.

weights were tested on dataset 1. Keeping the head's voting weight to a constant of 1, the weights of the frame and body are grid searched with a step size of 0.2 in the range

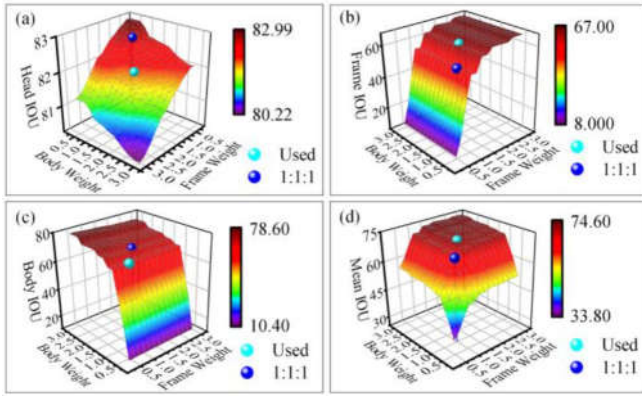


Fig. 9. Relationship between point cloud segmentation IOU and voting weights. The blue ball represents voting weights of 1: 1: 1; the cyan ball represents the weights calculated in this article. (a) Trend of head IOU with weight. (b) Trend of head frame IOU with weight. (c) Trend of body IOU with weight. (d) Trend of average IOU with weight.

of 0.4 to 3.0. The results are shown in Fig. 9, where Fig. 9a, 9b and 9c represent the relationship of the weight coefficient to the patient's head IOU, frame IOU and body IOU, respectively. Fig. 9d shows the relationship between the average IOU of all objects and the weight coefficient. The cyan ball in the figure represents the weights calculated by Equation (5), while the blue ball represents equal weights. During multiview voting, the voting weights of different objects greatly influence segmentation accuracy. Segmentation accuracy with respect to different objects can be improved by using appropriate weights, and the weight estimation calculated by Equation (5) obtains a relatively good segmentation result for each object.

IV. DISCUSSION

In this paper, a data-driven pipeline was explored to help stereotactic surgical robots understand unstructured surgical working scenes. We constructed a stereotactic surgery robot scene dataset and achieved object detection and segmentation from clinical 2D and 3D visual data. In each individual scene, both the prediction information based on the clinical big dataset and individual scene information, such as color and spatial distribution, were considered and effectively fused. This approach achieves clear benefits on both 2D and 3D data. We also proposed a personalized weight projection voting strategy that mitigates the unbalanced 3D segmentation performance problem caused by the varying 2D detection and segmentation accuracies for different objects.

Stereotactic surgical robots are an important type of surgical robot. As stereotactic surgical robots are being increasingly used in clinical fields, such as neurosurgery and orthopedics, how to make surgical robots more intelligent and improve surgical efficiency and safety has gradually become a new clinical concern. The ever-increasing quantity of surgical data for stereotactic robots also makes it possible to combine deep learning with stereotactic surgical robots. Real-time geometrical interpretation of the surgical scene is a fundamental aspect of autonomous surgical robotic systems. Geometrical information from surgical scenes is the first step needed to

develop advanced control techniques for medical robots. This paper presents a method enabling stereotactic robots to perceive and understand surgical scenes. To the best of our knowledge, we are the first to report on the use of computer vision and deep learning technology to help stereotactic surgical robots understand surgical scenes. Given the rapid development of computer vision and deep learning technology, there must be better solutions suitable for stereotactic surgical robot scene understanding—whether the scenes are 2D or 3D. Our method can provide a basic reference for future work.

As the object of this type of surgical operation, the patient's head is the most important target for stereotactic neurosurgical robots to understand. As shown in the results section, the method provides stable and accurate performances in patient head detection and segmentation at both the 2D- and 3D-scene levels; this is important in clinical practice.

In some cases, the head frame may collide with the robotic arm; thus, the head frame is another object that the robot needs to recognize and focus on. The head frame is a slender metal frame structure. Depending on the patient's positioning, different surgeries use different types of head frames. Even for the same type of operation, the head frames may have different spatial configurations. Therefore, the shape of the head frame varies substantially in various scenes. Accordingly, the results in Tables I and II show that the frame's detection and segmentation accuracy is lower than that of the head. This implies that a more suitable network or more clinical training data may be needed to understand these more complex objects. Moreover, because a factory 3D model of the head frame is available in advance, it is possible to consider incorporating the prior 3D model into the training in the future to achieve a better understanding of the head frame.

The patient's body usually occupies a large part of the scene but is always away from the robot. The patient's body is a noncritical target for which neurosurgical robots do not need to distinguish details. The test results show that the proposed model's detection and segmentation accuracy scores are relatively low for the body. However, the body is not the primary target. Therefore, while this situation requires further improvement, the current results are acceptable.

If robots can understand natural objects in the surgical environment, more solutions are possible for existing stereotactic surgical robot research. For example, in marker-free patient head tracking studies [40], complex filtering operations are often required to extract the patient's head-feature points from the background. However, by using the proposed 2D-scene-understanding method, the head-feature points can be easily extracted by the head mask. Moreover, in point cloud-based patient head registration, traditional solutions often assume that the patient's head point cloud is spherical-like; this paper tries to explore a data-driven pipeline where the head point cloud derives from unstructured environmental point clouds [31]. However, due to size differences and the imperfect shape of the patient's head, it is difficult to find a suitable spherical radius value that is neither too small to completely extract the head point cloud nor too large to avoid introducing point clouds of other objects around the head. As a comparison, our multiview projection voting method can directly

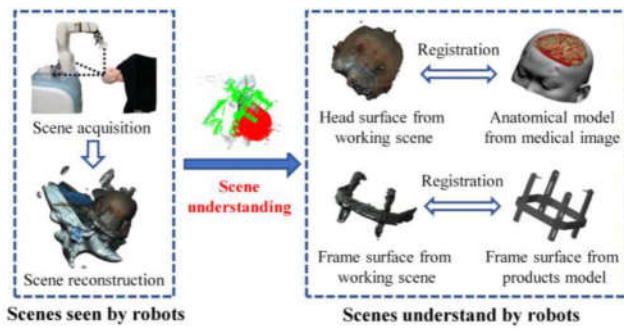


Fig. 10. 3D target object extraction combined with 3D scene understanding and prior model information.

extract the point cloud of the target object from the personalized 3D cloud point scene. Tables IV and V show that our method can obtain good extraction accuracy of the patient's head point cloud without requiring additional artificial estimation of the patient's head features. As shown in Fig. 10, the proposed 3D-scene-understanding method can be further combined with prior model information (such as information on an anatomical model from a medical image or a factory model), thereby allowing robots to obtain almost perfect 3D target object information and providing a map for registration or robot arm path planning tasks.

In our work, the deep learning method was partially adopted. For 2D-scene understanding, the traditional CRF and superpixel segmentation are introduced to fuse with the Mask R-CNN segmentation results. For 3D-scene understanding, the 3D region growing-based supervoxel algorithm was integrated with the multiview projection voting method. The results in Fig. 6 and Tables II-V show that a better performance is obtained by fusing the traditional image and point cloud processing methods with the neural network. The superpixels, CRF, and region growing method may not be the best choice for this problem and may even introduce increased time consumption, but the combination of existing interpretable traditional methods with clear models and physical meanings can achieve better performance and may alleviate the risks produced by unexplainable characteristics of neural networks [29]. For future applications in surgical robots, the method of combining data-driven neural networks and knowledge-driven mathematical models can be a technical route worth exploring.

In cases where the multiview projection voting method is used, 2D images objects that are segmented with high accuracy and have a high appearance frequency can reach an acceptable point cloud recognition voting threshold more easily, while objects that are segmented with low accuracy and have a low appearance frequency are more difficult to recognize. To alleviate this problem, we designed a personalized voting weighting mechanism to independently adjust the 3D scene segmentation performance for different target objects. The weight calculation method shown in Equation (5) can achieve a more balanced target segmentation performance between different target objects. Additionally, this method enables us to adjust the personalized weight according to different tasks.

For example, in the head point cloud-based registration, a more complete patient's head point cloud can be extracted by increasing the weight of the patient's head object (or decreasing the weight of the head frame and the patient's body object) so that we can obtain a head point cloud with a higher IOU, as shown in Fig. 9(a). However, Fig. 9(b) and 9(c) also show that the integrity of the patient's head frame and body will be reduced.

For stereotactic surgical robots, scene intelligence is a key part of perception, decision making, and action chains [41] and is the key to reaching a higher degree of automation. As the results of this study show, using neural networks to obtain the intelligence required for surgical robot perception by analyzing real clinical data appears very promising. However, in medical applications with strict safety requirements, ensuring stability and credibility is the most practical consideration for such methods. In future work, we will further improve and compare the methods in this article to achieve better performance and stability. The clinical training dataset will be further expanded, and more complex clinical scenario factors will be considered. In addition, for more intelligent surgical robots, the ability to understand not only the geometric information of the target object but also higher-level scene information needs to be studied. For example, how do robots perceive risks in surgery? How can risk information be effectively passed to the surgeon during surgery? All of these facets need to be further studied so that the robot can be truly aware of the surgical environment and become more intelligent, safe and easy to use.

V. CONCLUSION

The work described in this paper allows a stereotactic surgery robot to obtain the 2D/3D spatial position and semantic information of important clinical objects. The proposed method can stably and reliably extract head information from an unstructured surgical environment and achieve relatively precise detection and segmentation of the patient's head frame and body. Using the information provided by the proposed method, a stereotactic surgery robot will be able to automatically register medical images by identifying the point cloud of the patient's head and automatically designing a more optimized movement path based on the space occupied by the patient's head, the head frame, and other objects. These findings lay the foundation for developing stereotactic surgical robots with greater autonomy.

ACKNOWLEDGMENT

The author would like to thank Dr. Wenbo Liu from Sinovation Medical Technology Co., Ltd. (Beijing, China) for providing technical support for the research and the editors and anonymous reviewers who have given insightful comments.

REFERENCES

- [1] T. Haidegger, "Autonomy for surgical robots: Concepts and paradigms," *IEEE Trans. Med. Robot. Bionics*, vol. 1, no. 2, pp. 65–76, May 2019.
- [2] Y. Kassahun *et al.*, "Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 4, pp. 553–568, Apr. 2016.

- [3] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, "CAI4CAI: The rise of contextual artificial intelligence in computer-assisted interventions," *Proc. IEEE*, vol. 108, no. 1, pp. 198–214, Jan. 2020.
- [4] D. Wesierski and A. Jezierska, "Instrument detection and pose estimation with rigid part mixtures model in video-assisted surgeries," *Med. Image Anal.*, vol. 46, pp. 244–265, May 2018.
- [5] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Comput. Methods Programs Biomed.*, vol. 158, pp. 135–146, May 2018.
- [6] H. Nakawala, R. Bianchi, L. E. Pescatori, O. De Cobelli, G. Ferrigno, and E. De Momi, "'Deep-Onto' network for surgical workflow and context recognition," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, pp. 685–696, Apr. 2019.
- [7] C. Faria, W. Erlhagen, M. Rito, E. De Momi, G. Ferrigno, and E. Bicho, "Review of robotic technology for stereotactic neurosurgery," *IEEE Rev. Biomed. Eng.*, vol. 8, pp. 125–137, Apr. 2015. [Online]. Available: <https://ieeexplore.ieee.org/document/7098352/citations?tabFilter=papers#anchor-paper-citations-ieee>
- [8] S. Weber *et al.*, "Instrument flight to the inner ear," *Sci. Robot.*, vol. 2, no. 4, p. 12, Mar. 2017.
- [9] A. Ghasem, A. Sharma, D. N. Greif, M. Alam, and M. Al Maaieh, "The arrival of robotics in spine surgery: A review of the literature," *Spine*, vol. 43, no. 23, pp. 1670–1677, Dec. 2018.
- [10] E. De Momi and G. Ferrigno, "Robotic and artificial intelligence for key-hole neurosurgery: The ROBOCAST project, a multi-modal autonomous path planner," *Proc. Inst. Mech. Eng. H, J. Eng. Med.*, vol. 224, no. H5, pp. 715–727, 2010.
- [11] J. E. Bardram and N. Nørskov, *A Context-Aware Patient Safety System for the Operating Room*. New York, NY, USA: Assoc. Comput. Mach., 2008.
- [12] T. Beyl, P. Nicolai, M. D. Compartmenti, J. Raczkowski, E. De Momi, and H. Wörn, "Time-of-flight-assisted Kinect camera-based people detection for intuitive human robot cooperation in the surgical operating room," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 11, no. 7, pp. 1329–1345, Jul. 2016.
- [13] H. S. Yu *et al.*, "Methods and datasets on semantic segmentation: A review," *Neurocomputing*, vol. 304, pp. 82–103, Aug. 2018.
- [14] E. Z. Che, J. Jung, and M. J. Olsen, "Object recognition, segmentation, and classification of mobile laser scanning point clouds: A state of the art review," *Sensors*, vol. 19, no. 4, p. 810, Feb. 2019.
- [15] L. Wang *et al.*, "Multi-view fusion-based 3D object detection for robot indoor scene perception," *Sensors*, vol. 19, no. 19, p. 4092, Oct. 2019.
- [16] A. Haque, A. Milstein, and L. Fei-Fei, "Illuminating the dark spaces of healthcare with ambient intelligence," *Nature*, vol. 585, no. 7824, pp. 193–202, Sep. 2020.
- [17] S. Yeung *et al.*, "A computer vision system for deep learning-based detection of patient mobilization activities in the ICU," *NPJ Digit. Med.*, vol. 2, p. 5, Mar. 2019.
- [18] D. C. Birkhoff, A. S. H. M. van Dalen, and M. P. Schijven, "A review on the current applications of artificial intelligence in the operating room," *Surg. Innovat.*, vol. 28, no. 5, pp. 611–619, 2021.
- [19] L. R. Kennedy-Metz *et al.*, "Computer vision in the operating room: Opportunities and caveats," *IEEE Trans. Med. Robot. Bionics*, vol. 3, no. 1, pp. 2–10, Feb. 2021.
- [20] Z. Li, A. Shaban, J.-G. Simard, D. Rabindran, S. DiMaio, and O. Mohareri, "A robotic 3D perception system for operating room environment awareness," Mar. 2020, *arXiv:2003.09487*.
- [21] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: State of the art," *Int. J. Multimedia Inf. Retrieval*, vol. 9, no. 3, pp. 171–189, Sep. 2020.
- [22] K. M. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, 2017, pp. 2980–2988.
- [23] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT++: Better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 5, 2020, doi: [10.1109/TPAMI.2020.3014297](https://doi.org/10.1109/TPAMI.2020.3014297).
- [24] M. M. Rahman, Y. H. Tan, J. Xue, and K. Lu, "Notice of violation of IEEE publication principles: Recent advances in 3D object detection in the era of deep neural networks: A survey," *IEEE Trans. Image Process.*, vol. 29, pp. 2947–2962, 2020.
- [25] C. R. Qi, H. Su, K. C. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 77–85.
- [26] Y. Xie, J. Tian, and X. X. Zhu, "Linking points with labels in 3D: A review of point cloud semantic segmentation," *IEEE Geosci. Remote Sens. Mag.*, vol. 8, no. 4, pp. 38–59, Dec. 2020.
- [27] B. Li, "3D fully convolutional network for vehicle detection in point cloud," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Vancouver, BC, Canada, 2017, pp. 1513–1518.
- [28] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 945–953.
- [29] V. Monga, L. Yuelong, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," 2019, *arXiv:1912.10557*.
- [30] L. Li, J. Wu, H. Ding, and G. Wang, "A 'eye-in-body' integrated surgery robot system for stereotactic surgery," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 12, pp. 2123–2135, Jul. 2019.
- [31] F. L. Meng, F. W. Zhai, B. W. Zeng, H. Ding, and G. Z. Wang, "An automatic markerless registration method for neurosurgical robotics based on an optical camera," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 2, pp. 253–265, Feb. 2018.
- [32] B. W. Zeng, F. L. Meng, H. Ding, and G. Z. Wang, "A surgical robot with augmented reality visualization for stereoelectroencephalography electrode implantation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 8, pp. 1355–1368, Aug. 2017.
- [33] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 740–755.
- [34] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," Oct. 2012, *arXiv:1210.5644*.
- [35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
- [36] J. Papon, A. Abramov, M. Schoeler, and F. Wörgötter, "Voxel cloud connectivity segmentation—Supervoxels for point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2027–2034.
- [37] C. C. Wu, "Towards linear-time incremental structure from motion," in *Proc. Int. Conf. 3D Vis.*, Seattle, WA, USA, 2013, pp. 127–134.
- [38] C. C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 3057–3064.
- [39] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2010.
- [40] A. Z. Kyme, S. Se, S. R. Meikle, and R. R. Fulton, "Markerless motion estimation for motion-compensated clinical brain imaging," *Phys. Med. Biol.*, vol. 63, no. 10, May 2018, Art. no. 105018.
- [41] G. Z. Yang *et al.*, "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.*, vol. 2, no. 4, p. 2, Mar. 2017.