Article

Not peer-reviewed version

Toward Intraoperative Visual Intelligence: Real-Time Surgical Instrument Segmentation for Enhanced Surgical Monitoring

Mostafa Daneshgar Rahbar *, George Pappas , Nabih Jaber

Posted Date: 7 May 2024

doi: 10.20944/preprints202405.0300.v1

Keywords: Intraoperative surgery; monitoring surgical scene; convolutional neural network; U-Net; data augmentation; surgical tools segmentation; computer vision; image processing

Article

Toward Intraoperative Visual Intelligence: Real-Time Surgical Instrument Segmentation for Enhanced Surgical Monitoring

Mostafa Daneshgar Rahbar [1]*, George Pappas [2] and Nabih Jaber [3]

[1] Department of Electrical and Computer Engineering, Lawrence Technological University, Southfield, MI 48075, USA; mrahbar@ltu.edu
[2] Department of Electrical and Computer Engineering, Lawrence Technological University, Southfield, MI 48075, USA; gpappas@ltu.edu
[3] Department of Electrical and Computer Engineering, Lawrence Technological University, Southfield, MI 48075, USA; njaber@ltu.edu
* Correspondence: mrahbar@ltu.edu

Abstract: Background: Open surgery relies heavily on the surgeon's visual acuity and spatial awareness to track instruments within a dynamic and often cluttered surgical field. Methods: This system utilizes a head-mounted depth camera to monitor the surgical scene, providing both image data and depth information. Video captured from this camera is scaled down, compressed using MPEG, and transmitted to a high-performance workstation via RTSP (Real-Time Streaming Protocol), a reliable protocol designed for real-time media transmission. To segment surgical instruments, we utilize the Enhanced U-Net with Grid Mask (EUGNet) for its proven effectiveness in surgical tool segmentation. Results: For rigorous validation, the system's performance reliability, and accuracy are evaluated using prerecorded RGB-D surgical videos. This work demonstrates the potential of this system to improve situational awareness, surgical efficiency, and generate data-driven insights within the operating room. In a simulated surgical environment, the system achieved a high accuracy of 85.5% in identifying and segmenting surgical instruments. Furthermore, the wireless video transmission proved reliable with a latency of 200ms, suitable for real-time processing. Conclusions: These findings represent a promising step towards the development of assistive technologies with the potential to significantly enhance surgical practice.

Keywords: intraoperative surgery; monitoring surgical scene; convolutional neural network; U-Net; data augmentation; surgical tools segmentation; computer vision; image processing

1. Introduction

Traditionally, surgery relied solely on the surgeon's direct view for guidance. However, the past decade has witnessed a surge in automated analysis of surgical video data. These techniques offer surgeons various benefits, such as generating post-operative reports [1, 2], evaluating surgical skills for training purposes [3], or creating educational content. Furthermore, real-time video analysis holds promise for intraoperative communication with surgeons. This includes the development of automated warning or recommendation systems based on real-time recognition of surgical tasks, steps, or gestures [4-6]. Such systems could detect deviations from standard surgical procedures, potentially improving safety. However, a major challenge remains in interpreting surgical video data effectively: accurate detection of all surgical instruments. These instruments come in a wide variety of shapes and sizes, and often appear partially obscured within the surgical field. Many studies have focused on tackling this problem of surgical instrument detection. In the realm of minimally invasive surgery, particularly laparoscopy, technological advancements have revolutionized procedures. Laparoscopic surgeons rely on endoscopic cameras for viewing the surgical field, allowing for minimally invasive interventions. Similar technological progress can potentially improve

 doi:10.20944/preprints202405.0300.v1

visualization and coordination of intricate surgical maneuvers in traditional open surgeries, enhancing surgical precision and efficiency on standard video monitors.

Open surgery relies heavily on the surgeon's visual acuity and spatial awareness to track instruments within a dynamic and often cluttered surgical field. Maintaining focus and instrument identification can be challenging, especially during long procedures or intricate maneuvers. This project addresses this challenge by developing a novel visual intelligence system utilizing an optical approach for real-time surgical instrument tracking. The significance of this project lies in its ability to monitor the open surgery scene using a head-mounted depth camera. The system then segments surgical instruments in real-time from the captured video stream. This real-time segmentation offers several key advantages: **1) Enhanced Situational Awareness:** By segmenting instruments, the system can visually highlight them, aiding the surgeon in quickly identifying and tracking their location within the surgical field. This is particularly beneficial during minimally invasive procedures or situations with obscured views. **2) Improved Surgical Efficiency:** Real-time instrument tracking can potentially streamline surgical workflow by reducing the time spent searching for or confirming instrument location. **3) Potential for Data-Driven Insights:** Segmented instrument data can be further analyzed to provide valuable insights into surgical technique, instrument usage patterns, and potentially even identify potential errors or complications.

This project focuses on achieving real-time instrument segmentation through a system design that incorporates: **a) Head-Mounted Depth Camera:** This allows for a hands-free approach to capturing the surgical scene from the surgeon's perspective. **b) Wireless Video Transmission:** Real-time transmission of the captured video stream to a high-performance computer workstation enables the complex image processing required for segmentation. **c) Real-Time Instrument Segmentation:** The system segments surgical instruments from the video stream on the high-performance computer, providing immediate feedback to the surgeon. By achieving these goals, this project contributes to the advancement of intelligent surgical assistance systems, aiming to improve the overall safety, efficiency, and potentially the quality of open surgical procedures. The data collected by this system extends beyond immediate intraoperative safety. It can be used to establish safety volumes and define safe working zones for instruments, aiding in risk assessment [7]. It also can be utilized to develop smart glass augmentations and provide visual cues and warnings to improve surgical precision [8]. Last but not least, it can be used to enhance surgical training with Offering feedback on technique and movement patterns for inexperienced surgeons, facilitating faster learning [9, 10].

The importance of monitoring open surgery scenes and segmenting out surgical instruments is underscored by the need to enhance visualization and control during procedures. Hasan [11] and Hajj [12] both highlight the significance of segmenting and removing surgical instruments to improve the surgeon's view and facilitate automated instrument detection. Payandeh [13] and Panait [14] further emphasize the potential for image processing and video image enhancement to enhance surgical skill and facilitate finer maneuvers. However, the use of video images on standard monitors in open surgeries may lead to longer task performance [15]. Reiner [16] and Padoy [17] explore the potential of reproducing stereoscopic images and monitoring surgical procedures, respectively, to further improve visualization and workflow.

A range of studies have explored the use of advanced technologies in surgical settings. Islam [18] and Shvets [19] both developed deep learning-based systems for real-time instrument segmentation, with Islam's system outperforming existing algorithms. Fan [20] and Novotny [21] focused on 3D visualization and tracking, with Fan's system achieving a spatial position error of 0.38±0.92mm. Gering [22] and Dergachyova [23] integrated image fusion and interventional imaging, and proposed a data-driven method for surgical phase segmentation and recognition, respectively. Su [24] and Zhao [25] both developed real-time segmentation and tracking systems, with Su's algorithm achieving robust performance without GPU acceleration. These studies collectively demonstrate the potential of advanced technologies in enhancing surgical procedures.

Monitoring the surgical scene during open surgery using an optical approach is the main aim of this research. This work was to develop and validate a real-time visual intelligence system for tracking and segmenting surgical instruments during open surgery. Instrument Tracking Accurately

 doi:10.20944/preprints202405.0300.v1

tracking surgical instruments in the dynamic and sometimes chaotic environment of an operating room is one of most important underlying challenges of the research. Instrument Identification/Segmentation Distinguishing surgical instruments from other objects in the surgical field (hands, tissues, etc.) is another problem that this work tackled. Furthermore, real-time Monitoring Providing surgeons with updated information quickly, minimizing delays that could hinder the surgical process is another important challenge that this research tried to find a solution for.

This system utilizes a head-mounted depth camera to monitor the surgical scene, providing both image data and depth information. The surgeon's perspective facilitates a clear view of the surgical field. Video captured from this camera is scaled down, compressed using MPEG, and transmitted to a high-performance workstation via RTSP (Real-Time Streaming Protocol), a reliable protocol designed for real-time media transmission. The received video undergoes preprocessing, including optical flow filtering and other image processing techniques to enhance relevant features. To segment surgical instruments, we employ a convolutional neural network (CNN) approach. Specifically, we utilize the Enhanced U-Net with GridMask (EUGNet) [26] for its proven effectiveness in surgical tool segmentation.

For rigorous validation, the system's performance, reliability, and accuracy are evaluated using prerecorded RGB-D surgical videos. This work demonstrates the potential of this system to improve situational awareness, surgical efficiency, and generate data-driven insights within the operating room. In a simulated surgical environment, the system achieved a high accuracy of 85.5% in identifying and segmenting surgical instruments. Furthermore, the wireless video transmission proved reliable with a latency of 200ms, suitable for real-time processing. These findings represent a promising step towards the development of assistive technologies with the potential to significantly enhance surgical practice.

2. Materials and Methods

This system utilizes a head-mounted depth camera to monitor the surgical scene, providing both image data and depth information. The surgeon's perspective facilitates a clear view of the surgical field. Video captured from this camera is scaled down, compressed using MPEG, and transmitted to a high-performance workstation via RTSP (Real-Time Streaming Protocol), a reliable protocol designed for real-time media transmission. The received video undergoes preprocessing, including optical flow filtering and other image processing techniques to enhance relevant features. To segment surgical instruments, we employ a convolutional neural network (CNN) approach. Specifically, we utilize the Enhanced U-Net with GridMask (EUGNet) [26] for its proven effectiveness in surgical tool segmentation.

2.1. Head-Mounted Depth Camera

The use of RGB-D depth cameras, such as the Intel RealSense series, offers significant advantages for real-time intraoperative surgery monitoring. The RGB component provides traditional visual information for instrument identification, while the depth component enables a more comprehensive understanding of the surgical field's 3D structure. This depth information aids in accurately tracking instrument locations, even in complex or cluttered environments. Additionally, depth data can enhance instrument segmentation algorithms, improving their ability to distinguish surgical tools from the background and other objects with similar visual appearances. The integration of an RGB-D RealSense camera within a surgical monitoring system holds the potential to enhance procedural safety, efficiency, and the collection of valuable intraoperative data.

RGB-D real sense technology can be useful for intraoperative surgery monitoring by providing real-time 3D visualization and tracking of surgical instruments and anatomical structures [27]. It utilizes depth sensors like the Intel RealSense to capture color and depth data, enabling augmented reality overlays and enhanced navigation during procedures [28]. This real-time imaging modality complements traditional intraoperative imaging techniques like ultrasound and MRI, offering additional spatial awareness and guidance [29]. RGB-D sensing can track respiratory motion and

 doi:10.20944/preprints202405.0300.v1

deformation of soft tissues, allowing for more precise targeting and compensation during interventions [27]. The combination of color and depth data provides enhanced visualization of the surgical field, aiding in instrument navigation and identification of critical structures [28]. Overall, RGB-D real sense technology shows promise as an intraoperative imaging modality, offering real-time 3D guidance and augmented reality capabilities to improve surgical precision and safety.

Head-mounted RGB-D sensors like RealSense can be useful for intraoperative monitoring during posterior skull base surgery [30]. They allow real-time visualization of surgical instruments and anatomical structures, providing enhanced guidance to the surgeon [31]. The system uniquely employs a head-mounted RGB-D camera, positioning it directly at the surgeon's perspective. This setup provides a wider field of view and generates two crucial data streams: an RGB frame for visual context and a point cloud for 3D spatial analysis.

2.2. UP Squared Board

For data transfer, the system employs a UP Squared board. This compact computer acts as a wireless bridge, receiving the captured RGB video and point cloud data stream from the surgeon-mounted RealSense camera. The UP Squared board then transmits this data wirelessly to a high-performance workstation for real-time analysis. The use of the UP Squared board for wireless data transfer aligns with previous research on real-time remote monitoring systems. [32, 33] both highlight the use of wireless communication for data acquisition and transmission, with Yan 2010 specifically focusing on a low-cost wireless bridge. The real-time aspect of the system is further supported by [34], who presents a wireless solution for data acquisition in a real-time environment. The system's ability to capture and calibrate video data is also in line with [35] on a low-cost hardware platform for video capture. The use of a single-chip microcomputer for network data transfer, as discussed by [35], could potentially enhance the efficiency of the system. Lastly, the system's potential for high-efficiency video transmission is supported by [36] on a compressed sensing-based video transmission system.

2.3. High Performance Work Station

Video captured from Real Sense camera is preprocessed and transmitted to High performance workstation. It is a computer with an 13th Generation Intel Core™ i9-13900KF Processor (E-cores up to 4.30 GHz P-cores up to 5.40 GHz) CPU and a NVIDIA GeForce RTX™ 4080 16GB GDDR6X GPU. The inference time was calculated including data transfers from CPU to GPU and back and averaged across 1000 inferences. scaled down, compressed using MPEG, and transmitted to a high-performance workstation via RTSP (Real-Time Streaming Protocol), a reliable protocol designed for real-time media transmission. The received video undergoes preprocessing, including optical flow filtering and other image processing techniques to enhance relevant features. As It is mentioned earlier, to segment surgical instruments, we employ a convolutional neural network (CNN) approach. Specifically, we utilize the Enhanced U-Net with GridMask (EUGNet) [26] for its proven effectiveness in surgical tool segmentation. **Error! Reference source not found.** summarize the different components of designed intraoperative visual intelligence system.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2024 doi:10.20944/preprints202405.0300.v1

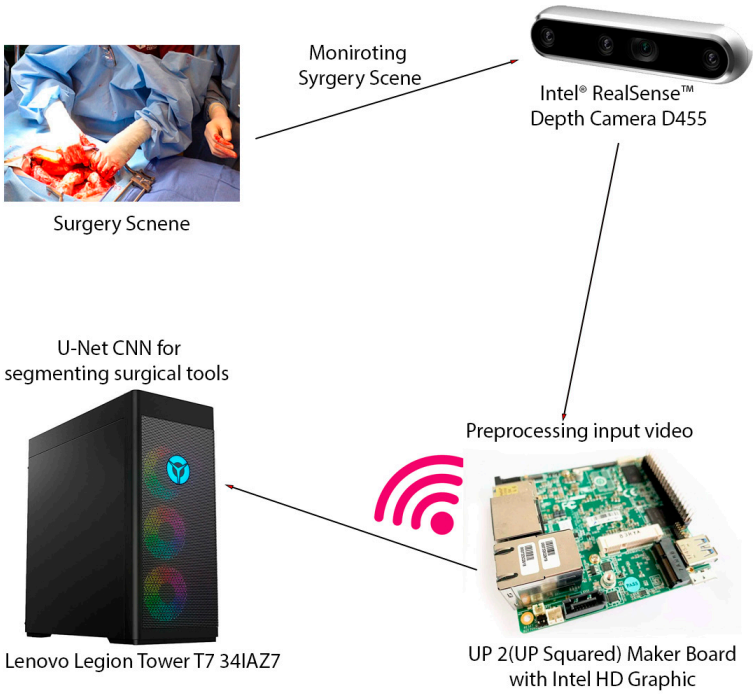

Figure 1. Hardware components of intraoperative visual intelligence system.

2.4. Middleware Framework

ROS is a middleware framework designed for robot software development. It facilitates communication between different robot components and offers tools for common robotic tasks like sensor data processing and motion control. Our system is leveraging specific ROS components to enhance certain aspects of your surgical instrument tracking project. ROS is used for data acquisition and visualization. Since Real Sense is ROS-compatible cameras, we leverage existing ROS drivers to streamline data acquisition and camera integration: If you're using. ROS provides standardized message formats that simplify integrating video and depth data. ROS is also utilized for Visualization. RViz, a powerful ROS visualization tool, could be used to display camera data alongside overlays indicating segmented instruments. This could provide real-time feedback during development or for intraoperative monitoring. Modular system design is another reason that ROS is selected as Middleware Framework. Its Node-Based Architecture is employed to structure parts of your system as ROS nodes. This promotes modularity, where image processing, instrument segmentation, and data transmission components can communicate via ROS topics and services. We find existing ROS packages for image processing or computer vision tasks that could integrate seamlessly into your system. Gazebo, a simulator often used with ROS, is used to create simulated surgical environments. This helped to evaluate algorithms or generate synthetic training data. Because of its extensive libraries and tools, ROS is excellent for rapid prototyping of algorithms for tracking or segmentation.

To create a ROS (Robot Operating System) publish-subscribe diagram for your described system involving the segmentation of surgical tools and movement analysis using a head-mounted depth camera, I'll outline the different nodes and topics involved. Here's a structured plan for the nodes and the information flow in the system:

1. **Camera Node**
   - **Purpose:** Captures video and depth data from a head-mounted depth camera.
   - **Publishes:**
     - **Topic:** /camera/image_raw (Image data)
     - **Topic:** /camera/depth_raw (Depth data)
2. **Video Processing Node**
   - **Purpose:** Handles the reception, scaling down, and compression of video data.

 doi:10.20944/preprints202405.0300.v1

- **Subscribes:**
  - **Topic:** /camera/image_raw
  - **Processes:** Scales down and compresses video using MPEG, transmits it via RTSP.
- **Publishes:**
  - **Topic:** /video/compressed

3. **Preprocessing Node**
   - **Purpose:** Processes the compressed video to prepare for segmentation.
   - **Subscribes:**
     - Topic: /video/compressed
   - **Processes:** Applies optical flow filtering and other image processing techniques.
   - Publishes:
     - Topic: /video/processed
4. **Segmentation Node**
   - **Purpose:** Segments surgical instruments from the video.
   - **Subscribes:**
     - Topic: /video/processed
   - **Processes:** Uses a convolutional neural network (EUGNet with GridMask) for segmentation.
   - **Publishes:**
     - Topic: /video/segmented_tools
5. **Movement Analysis Node**
   - **Purpose:** Analyzes the movement of segmented tools.
   - **Subscribes:**
     - **Topic:** /video/segmented_tools
   - **Processes:** Conducts movement analysis.
   - **Publishes:**
     - **Topic:** /tools/movement_analysis
6. **Visualization Node**
   - **Purpose:** Visualizes the segmented tools and their movement analysis for monitoring and further analysis.
   - **Subscribes:**
     - **Topic:** /video/segmented_tools
     - **Topic:** /tools/movement_analysis

2.5. Surgical Instrument Segmentation Network

Enhanced U-Net with GridMask (EUGNet), which incorporates GridMask augmentation to address U-Net's limitations and is proposed in previous work [26] is utilized to do surgical instrument segmentation. EUGNet features a deep contextual encoder, residual connections, class-balancing loss, adaptive feature fusion, GridMask augmentation module, efficient implementation, and multi-modal fusion. These innovations enhance segmentation accuracy and robustness, making it well-suited for medical image analysis. The GridMask algorithm, designed for improved pixel elimination, demonstrates its effectiveness in enhancing model adaptability to occlusions and local features, crucial in dynamic surgical settings. A comprehensive dataset of robotic surgical scenarios and instruments rigorously evaluates the framework's robustness. Employing the U-Net architecture as a baseline, the use of GridMask as a data augmentation technique significantly improves both segmentation accuracy and inference speed. These results highlight GridMask's potential as a valuable tool for real-time instrument-tissue segmentation in robotic surgery. **Error! Reference source not found.** depicts Visual Representation of EUGNet.

 doi:10.20944/preprints202405.0300.v1

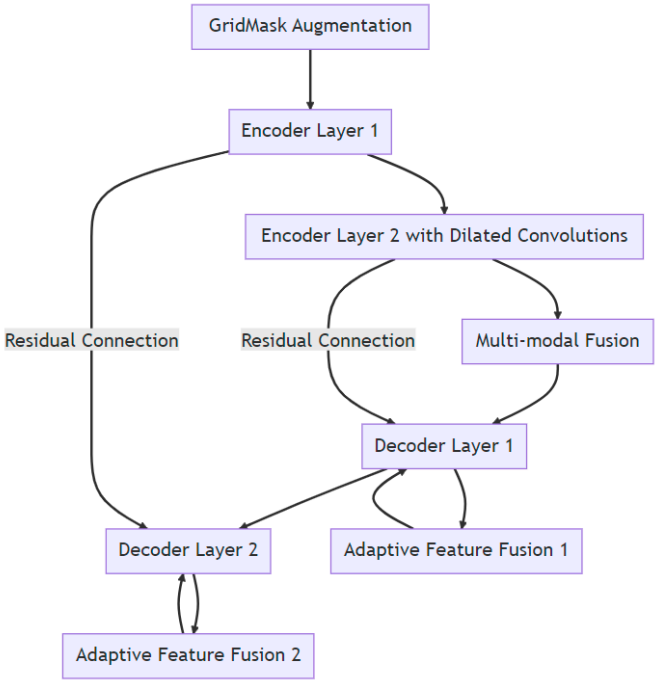

Figure 2. Visual Representation of EUGNet.

2.5. Data Collection for Algorithm Evaluation

To demonstrate the robustness and generalization ability of the proposed robotic instrument segmentation framework, we employed a dataset featuring diverse surgical scenarios and instruments. This dataset included the Da Vinci Robotic (DVR) Dataset [37] with four 45-second ex vivo training videos and six test videos (ranging from 15 to 60 seconds) showcasing two articulated instruments. Ground truth masks were automatically generated with manual correction, and videos were recorded using the dVRK Open Source Platform [38]. Additionally, we utilized open-source videos from the U.S. National Library of Medicine [39] depicting procedures like midline lobectomy, right superior line lobectomy, thoracotomy, thoracoscopic lung surgery, and prostatectomy with occurrences of "splash-like" bleeding. Also, datasets from the EndoVis challenge [40] for comprehensive evaluation . Finally, the Endoscapes Dataset [41] provided 201 laparoscopic videos richly annotated for scene segmentation, object detection, and critical view of safety assessment. All videos across these datasets had a resolution of 720x567 and ran at 25 frames per second.

2.6. Baseline Method and Evaluation Protocol

U-Net is a popular choice for medical image analysis, particularly for segmenting surgical instruments. As a baseline for comparison, we employed a state-of-the-art U-Net architecture known for its effectiveness in segmenting robotic surgical tools. This choice leverages the U-Net's well-established capabilities for this task. To improve the model's ability to generalize to unseen data during training, we utilized random image selection. We prioritized comparing our proposed architecture to this baseline over achieving the highest possible segmentation scores. Additionally, GridMask data augmentation was applied to further enhance the model's robustness. To avoid potential biases from the source dataset being introduced during transfer learning, we opted to train the model from scratch. In our experiments, the cyclical learning rate (CLR) bounds for the U-Net network are set to (1e-4; 1e-2). The quantitative metrics of choice to evaluate the predicted segmentations are mean intersection over union (mean IoU) and mean Dice similarity coefficient (mean DSC):

$$\overline{IoU}(\hat{y}, y) = \frac{1}{K}\sum_{k=1}^{K}\frac{TP_k}{TP_k + FP_k + NF_k} \quad (1)$$

 doi:10.20944/preprints202405.0300.v1

$$\overline{DSC}(\hat{y}, y) = \frac{1}{K}\sum_{k=1}^{K}\frac{2TP_k}{2TP_k + FP_k + NF_k} \quad (2)$$

where K = 2 (k = 0 background, k = 1 foreground), and TP_k, FP_k, and NF_k represent true positives, false positives, and false negatives for class k, respectively.

All networks were trained and tested (including inference times) on a computer with a 13th generation Intel Core™ i9-13900KF processor (E-cores up to 4.30 GHz and P-cores up to 5.40 GHz) CPU and a NVIDIA GeForce RTX™ 4080 16GB GDDR6X GPU. The inference time was calculated, including data transfers from CPU to GPU and back, and averaged across 1000 inferences.

The Intel RealSense D455 is an advanced stereo depth camera that extends the capabilities of its predecessors by providing greater accuracy and longer range sensing. Equipped with two high-resolution depth sensors and a RGB sensor, the D455 offers a depth range of up to 20 meters, making it suitable for a variety of applications, from robotics and augmented reality to more complex scenarios like gesture recognition and people tracking. Its improved accuracy and wider baseline of 95 mm between the depth sensors enhance depth perception and reduce blind spots, thus providing more precise 3D imaging.

One of the standout features of the RealSense D455 is its built-in IMU (Inertial Measurement Unit), which provides additional data points about device orientation and movement, enhancing the depth data with spatial awareness. This feature is particularly valuable for mobile applications, where understanding the device's position and orientation in space is crucial. The D455 is designed to be plug-and-play, supporting both Windows and Linux environments, and integrates seamlessly with the Intel RealSense SDK 2.0, which offers a rich set of libraries and APIs to expedite development and integration. Whether used for creating interactive experiences, developing navigation systems for drones and robots, or for enhanced computer vision in complex environments, the Intel RealSense D455 provides developers and engineers with a powerful tool to bring depth-sensing capabilities to their projects.

The Intel Up Squared (or UP²) board is a compact and powerful single-board computer designed for a variety of applications in the embedded computing, Internet of Things (IoT), and edge computing spaces. It represents an evolution in the Up board series, offering significantly enhanced processing power, flexibility, and connectivity compared to its predecessors.

The UP Squared board features Intel Apollo Lake processors, including Intel Celeron Intel® Pentium® J6426. This processor provide a balance between performance and power efficiency, making the board suitable for demanding tasks that also require low power consumption. It has 16GB DDR4 RAM and 64GB of eMMC storage, providing ample space and speed for various applications. Additionally, there is support for an M.2 SSD, enhancing its capabilities for storage-intensive applications. The board includes multiple USB ports (USB 3.0 and USB 2.0), Gigabit Ethernet, HDMI, and DisplayPort, facilitating a wide range of connectivity options for peripherals and displays. It also supports wireless connections via an M.2 slot that can host WiFi and Bluetooth modules. Due to its powerful features and multiple connectivity options, UP Squared is suitable for a wide range of applications such as digital signage, kiosks, IoT gateways, smart home devices, edge computing devices, and even as a development platform for AI and machine learning projects.

Intel® Wi-Fi 6 AX210 module is providing a future-proof option that supports the latest WiFi 6 technology. It Supports 802.11ax technology, which can significantly improve throughput and capacity, particularly in dense environments. It also offers Bluetooth 5.1 for extended range and capabilities. It is good for high-performance applications requiring the highest data rates and improved efficiency, such as streaming high-definition video, gaming, or handling multiple wireless devices simultaneously[42].

3. Results

In this study, ROS 2 was deployed on an UP Squared board operating under a Linux Ubuntu environment to establish a sophisticated video processing framework. A custom video publisher node written in C++ was developed to interface with an Intel RealSense camera, which captures video

 doi:10.20944/preprints202405.0300.v1

data in real-time. This node is responsible for the initial preprocessing and compression of the video data before publishing it to a designated ROS topic named realsense_compressed_image. A corresponding subscriber node, executed on a high-performance computing system as previously described, subscribes to the realsense_compressed_image topic. Both machines are configured within the same DDS (Data Distribution Service) domain, facilitating seamless intra-domain communication. This setup ensures that video data is efficiently captured, processed, and transmitted between the nodes without significant latency or loss of data integrity. **Error! Reference source not found.** in the manuscript illustrates the use of rqt_graph, a graphical tool provided by ROS for visualizing the computation graph of the system. rqt_graph effectively demonstrates the active nodes and their interconnections through topics within the ROS environment. This visualization is generated subsequent to the activation of the aforementioned nodes on both computers, providing a clear and intuitive representation of the dynamic interactions and data flow within the network. The integration of ROS 2 with the UP Squared board and RealSense camera showcases a robust platform for real-time video processing applications, highlighting the system's scalability and flexibility in handling complex computational tasks across distributed nodes. The wireless video transmission proved reliable with a latency of 200ms, suitable for real-time processing.

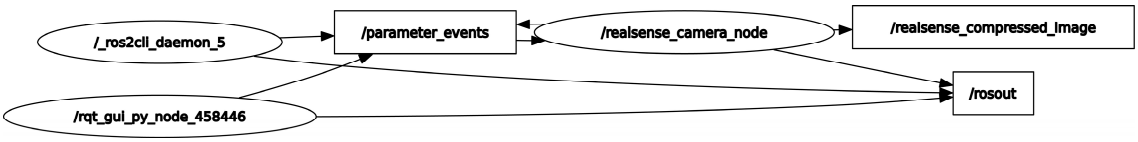

Figure 3. rqt_graph generated ROS communication.

Our proposed data augmentation technique, GridMask, significantly improved the performance of the U-Net architecture on the EndoVis testing set. Metrics including balanced accuracy (foreground), mean intersection over Union (IoU), and mean Dice Similarity Coefficient (DSC) all demonstrated substantial gains (specific values shown in see **Error! Reference source not found.**). Furthermore, GridMask drastically reduced inference time from 0.163ms for the baseline U-Net to 0.097ms for the U-Net with GridMask. This reduction facilitates real-time instrument-tissue segmentation, achievable at approximately 29 frames per second. Qualitative analysis (visual results not shown here) reveals that our method using GridMask achieves better adherence to the boundaries of left-handed surgical instruments.

Using GridMask data augmentation alongside the U-Net architecture, we achieved substantial improvements on the Endoscape dataset [41]. Metrics such as balanced accuracy, IoU, and mean DSC showed significant gains (see **Error! Reference source not found.** for specific values). Notably, GridMask drastically reduced inference time to 0.097ms, making the approach suitable for real-time instrument-tissue segmentation at approximately 29 fps. Qualitative analysis (see **Error! Reference source not found.**) demonstrates the enhanced adherence to tool boundaries when using GridMask, especially for left-handed surgical instruments.

Table 1. Quantitative results for segmentation of non-rigid robotic instruments in testing set videos. IoU stands for intersection over union, and DSC for Dice similarity coefficient. The means are performed over classes, and the results presented are averaged across testing frames.

Network	Inference Time (ms/fps)	Balanced Accuracy (fg.)	Mean IoU	Mean DSC
EUGNet with EndoVis Dataset	30.2/25.2	89.3%	84.6%	85.5%
EUGNet with Endoscape Dataset	31.7/26.7	84.3%	82.6%	81.5%

 doi:10.20944/preprints202405.0300.v1

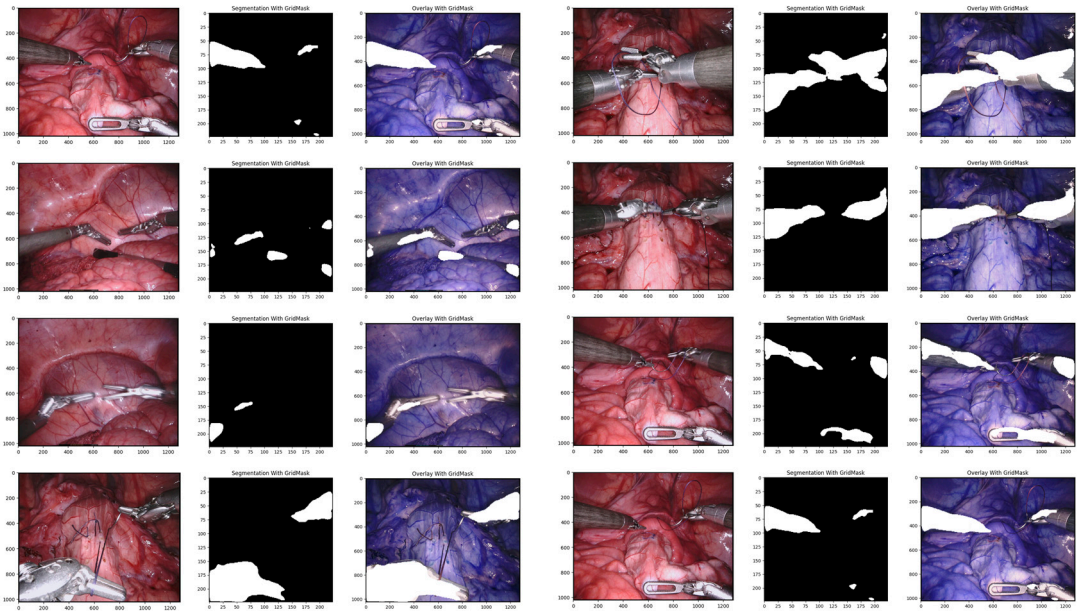

Figure 4. Qualitative comparison of our proposed convolutional architectures with GridMask data augmentation.

The training loss curve in **Error! Reference source not found.** demonstrates a stable decline over 40 epochs when using GridMask, indicating good model generalization and minimal overfitting. This suggests the model effectively learns from the training data without memorizing specific details. Furthermore, accuracy metrics fall within the "excellent" range [43], implying the model's predictions are highly reliable. The Dice coefficient, a crucial metric for segmentation tasks, also exhibits strong performance with GridMask, falling within the "excellent" range [44]. Notably, the consistently high Dice coefficient indicates a superior overlap between the model's predicted segmentation masks and the ground truth labels. Collectively, these findings suggest that GridMask data augmentation significantly improves the U-Net's ability to learn robust and accurate segmentation capabilities.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 May 2024 doi:10.20944/preprints202405.0300.v1

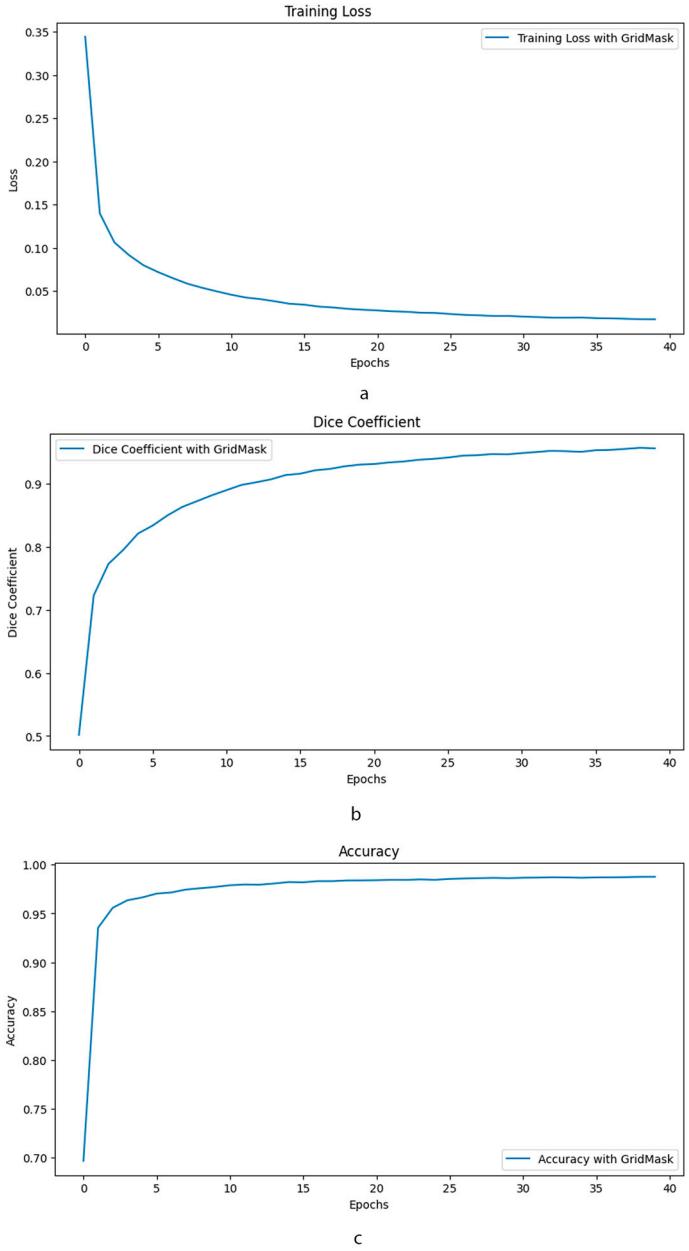

Figure 5. Performance metrics of enhanced U-Net with GridMask data augmentation. This composite image showcases three key performance indicators—training loss (**a**), Dice coefficient (**b**), and accuracy (**c**)—over 40 epochs, comparing the outcomes of an enhanced U-Net.

4. Discussion

A range of studies have explored the use of head-mounted cameras for real-time monitoring in the operating room. [45] and [46] both highlight the potential of smartphone and GoPro cameras, respectively, for capturing high-quality intraoperative footage. [47] provides an overview of the advantages and disadvantages of different video capture methods, including head-mounted cameras. Based on these surveys this technology is still a "work-in-progress" and when it comes to its applicability in the operating room, and it requires further fine-tuning to optimize its utility. Our proposed method is successfully utilized to monitor and captur higher quality video to monitor the surgical scene. It also, provide additional piece of information about the depth of surgical scene. As one possible extension of this work, the depth information can be converted to the point cloud object and transmit to the high-performance system to provide the information for refine segmentation of surgical instrument. Instance segmentation of point cloud is extremely important since the quality of

 doi:10.20944/preprints202405.0300.v1

segmentation will affect the performance of subsequent algorithms. The point cloud captured by RGB-D sensor can go over the instance segmentation process by applying the deep learning method YOLACT++ to instance segment the color image first and then matching the instance information with the point cloud [48].

This work presents a surgical instrument segmentation system that leverages a distributed ROS 2 architecture for real-time data transmission and processing. The system employs a head-mounted depth camera at the surgical site, capturing high-quality RGB video along with corresponding depth information. This data is transmitted, potentially wirelessly using a reliable protocol like RTSP (Real-Time Streaming Protocol), to a high-performance workstation for real-time instrument segmentation analysis. ROS 2 facilitates communication between the camera and the workstation, enabling a modular and scalable architecture for efficient data handling.

In comparison to our previous results in [26], downscaling captured video at the surgeon's site presents a potential reduction in the accuracy of U-Net surgical instrument segmentation. While reduced video resolution improves transmission efficiency and potentially processing speed, which are essential for real-time applications, it also results in a loss of fine visual details. This loss of detail can hinder the model's ability to precisely delineate tool boundaries, especially for small or thin instruments. Overlapping instruments may become harder to distinguish, and textural details useful for classification could be lost. While U-Net's skip connections offer some resilience to downscaling, excessive resolution reduction can limit their effectiveness. Understanding and mitigating these impacts is important. Experimentation is necessary to find the optimal balance between acceptable downscaling levels and segmentation accuracy, considering the specific surgical tasks of your project.

5. Conclusions

In this research, we have demonstrated a novel intraoperative visual intelligence system that enhances surgical monitoring by providing real-time segmentation of surgical instruments. This advancement leverages cutting-edge technologies, including head-mounted depth cameras and convolutional neural networks, to significantly improve the visibility and tracking of surgical tools, thereby enhancing surgical precision and safety. The system's ability to provide real-time feedback to surgeons is a critical development, particularly in complex surgeries where visibility is compromised. Our results indicate a high accuracy in instrument identification and segmentation, demonstrating the system's potential to not only support surgeons in real-time but also to serve as a valuable training tool for surgical education.

Looking forward, the integration of such technologies promises to revolutionize the operating room, reducing surgical errors and improving patient outcomes. This work lays the groundwork for further innovations in surgical procedures and offers a glimpse into the future of automated and enhanced surgical environments. Further studies and refinements will likely focus on optimizing the system's accuracy and responsiveness, expanding its application to various surgical contexts, and integrating deeper learning and adaptive algorithms to handle an even broader array of surgical instruments and conditions.

Author Contributions: Conceptualization, M.D.R. and G.P.; methodology, M.D.R., G.P., and N.J..; software, M.D.R.; validation, M.D.R.; formal analysis, M.D.R., G.P., and N.J.; investigation, M.D.R., G.P., and N.J.; resources, M.D.R.; data curation, M.D.R.; writing—original draft preparation, M.D.R., G.P., and N.J.; writing—review and editing, M.D.R., G.P., and N.J; visualization, M.D.R.; supervision, M.D.R., G.P., and N.J; project administration, M.D.R.; funding acquisition, M.D.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported with funding from the College of Engineering Research Seed Grant Program at Lawrence Technological University.

Data Availability Statement: The data presented in this study are available at https://universe.roboflow.com/models/instance-segmentation and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6462551/figure/vid/.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Lalys, F., et al., *A framework for the recognition of high-level surgical tasks from video images for cataract surgeries.* IEEE Transactions on Biomedical Engineering, 2011. **59**(4): p. 966-976.
2. Stanek, S.R., et al., *Automatic real-time detection of endoscopic procedures using temporal features.* Computer methods and programs in biomedicine, 2012. **108**(2): p. 524-535.
3. André, B., et al., *Learning semantic and visual similarity for endomicroscopy video retrieval.* IEEE Transactions on Medical Imaging, 2012. **31**(6): p. 1276-1288.
4. Quellec, G., et al., *Real-time segmentation and recognition of surgical tasks in cataract surgery videos.* IEEE transactions on medical imaging, 2014. **33**(12): p. 2352-2360.
5. Charriere, K., et al. *Automated surgical step recognition in normalized cataract surgery videos*. in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014. IEEE.
6. Quellec, G., et al., *Real-time task recognition in cataract surgery videos using adaptive spatiotemporal polynomials.* IEEE transactions on medical imaging, 2014. **34**(4): p. 877-887.
7. Glaser, B., S. Dänzer, and T. Neumuth, *Intra-operative surgical instrument usage detection on a multi-sensor table.* International journal of computer assisted radiology and surgery, 2015. **10**: p. 351-362.
8. Tsubosaka, M., et al., *Additional visualization via smart glasses improves accuracy of wire insertion in fracture surgery.* Surgical Innovation, 2017. **24**(6): p. 611-615.
9. Islam, G., B. Li, and K. Kahol. *Developing a real-time low-cost system for surgical skill training and assessment*. in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2013. IEEE.
10. Pinzon, D., et al., *Skill learning from kinesthetic feedback.* The American Journal of Surgery, 2017. **214**(4): p. 721-725.
11. Hasan, S.K., R.A. Simon, and C.A. Linte. *Segmentation and removal of surgical instruments for background scene visualization from endoscopic/laparoscopic video*. in *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. 2021. SPIE.
12. Hajj, H.A., et al., *Coarse-to-fine surgical instrument detection for cataract surgery monitoring.* arXiv preprint arXiv:1609.05619, 2016.
13. Payandeh, S., J. Hsu, and P. Doris, *Toward the design of a novel surgeon-computer interface using image processing of surgical tools in minimally invasive surgery.* International Journal of Medical Engineering and Informatics, 2012. **4**(1): p. 1-24.
14. Panait, L., et al., *Surgical skill facilitation in videoscopic open surgery.* Journal of Laparoendoscopic & Advanced Surgical Techniques, 2003. **13**(6): p. 387-395.
15. Mohamed, A., et al., *Skill performance in open videoscopic surgery.* Surgical Endoscopy And Other Interventional Techniques, 2006. **20**: p. 1281-1285.
16. Reiner, J., *Possibilities for reproducing stereoscopic images on monitors in relation to the surgical microscope.* Klinische Monatsblatter fur Augenheilkunde, 1990. **196**(1): p. 51-53.
17. Padoy, N., *Workflow and activity modeling for monitoring surgical procedures*. 2010, Université Henri Poincaré-Nancy 1; Technische Universität München.
18. Islam, M., et al., *Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning.* IEEE Robotics and Automation Letters, 2019. **4**(2): p. 2188-2195.
19. Shvets, A.A., et al. *Automatic instrument segmentation in robot-assisted surgery using deep learning*. in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. 2018. IEEE.
20. Fan, Z., et al., *3D interactive surgical visualization system using mobile spatial information acquisition and autostereoscopic display.* Journal of biomedical informatics, 2017. **71**: p. 154-164.
21. Novotny, P.M., et al. *Real-time visual servoing of a robot using three-dimensional ultrasound*. in *Proceedings 2007 IEEE international conference on robotics and automation*. 2007. IEEE.
22. Gering, D.T., et al. *An integrated visualization system for surgical planning and guidance using image fusion and interventional imaging*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI'99: Second International Conference, Cambridge, UK, September 19-22, 1999. Proceedings 2*. 1999. Springer.
23. Dergachyova, O., et al., *Automatic data-driven real-time segmentation and recognition of surgical workflow.* International journal of computer assisted radiology and surgery, 2016. **11**: p. 1081-1089.
24. Su, Y.-H., K. Huang, and B. Hannaford. *Real-time vision-based surgical tool segmentation with robot kinematics prior*. in *2018 International Symposium on Medical Robotics (ISMR)*. 2018. IEEE.
25. Zhao, Z., et al., *Real-time tracking of surgical instruments based on spatio-temporal context and deep learning.* Computer Assisted Surgery, 2019. **24**(sup1): p. 20-29.
26. Daneshgar Rahbar, M. and S.Z. Mousavi Mojab, *Enhanced U-Net with GridMask (EUGNet): A Novel Approach for Robotic Surgical Tool Segmentation.* Journal of Imaging, 2023. **9**(12): p. 282.
27. Özbek, Y., Z. Bárdosi, and W. Freysinger, *respiTrack: patient-specific real-time respiratory tumor motion prediction using magnetic tracking.* International Journal of Computer Assisted Radiology and Surgery, 2020. **15**(6): p. 953-962.
28. Shamov, T., et al., *Ultrasound-based neuronavigation and spinal cord tumour surgery-marriage of convenience or notified incompatibility?* Turkish Neurosurgery, 2013. **23**(3).

29. Tokuda, J., et al. *New 4-D imaging for real-time intraoperative MRI: Adaptive 4-D scan*. in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006: 9th International Conference, Copenhagen, Denmark, October 1-6, 2006. Proceedings, Part I 9*. 2006. Springer.
30. Kartush, J.M., et al., *Intraoperative cranial nerve monitoring during posterior skull base surgery.* Skull Base Surgery, 1991. **1**(02): p. 85-92.
31. Dick, A.J., et al., *Invasive human magnetic resonance imaging: feasibility during revascularization in a combined XMR suite.* Catheterization and cardiovascular interventions, 2005. **64**(3): p. 265-274.
32. Velásquez-Aguilar, J., et al. *Multi-channel data acquisition and wireless communication FPGA-based system, to real-time remote monitoring*. in *2017 International Conference on Mechatronics, Electronics and Automotive Engineering (ICMEAE)*. 2017. IEEE.
33. Linderman, L.E., K.A. Mechitov, and B.F. Spencer Jr, *Real-time wireless data acquisition for structural health monitoring and control.* Newmark Structural Engineering Laboratory Report Series 029, 2011.
34. Shah, D. and U.D. Dalal. *Wireless data assistance in real time environment using DSP processor*. in *Proceedings of the International Conference & Workshop on Emerging Trends in Technology*. 2011.
35. Chen, W. and X. Huang. *The Design and Application of Embedded Processor Based on Single Chip Microcomputer in Network Laboratory*. in *Proceedings of the 2017 International Conference on E-Society, E-Education and E-Technology*. 2017.
36. Zheng, S., et al., *A high-efficiency compressed sensing-based terminal-to-cloud video transmission system.* IEEE transactions on multimedia, 2019. **21**(8): p. 1905-1920.
37. Pakhomov, D., et al. *Deep residual learning for instrument segmentation in robotic surgery*. in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10*. 2019. Springer.
38. Kazanzides, P., et al. *An open-source research kit for the da Vinci® Surgical System*. in *2014 IEEE international conference on robotics and automation (ICRA)*. 2014. IEEE.
39. Novellis, P., et al., *Management of robotic bleeding complications.* Annals of cardiothoracic surgery, 2019. **8**(2): p. 292.
40. Roß, T., et al., *Comparative validation of multi-instance instrument segmentation in endoscopy: results of the ROBUST-MIS 2019 challenge.* Medical image analysis, 2021. **70**: p. 101920.
41. Murali, A., et al., *The Endoscapes Dataset for Surgical Scene Segmentation, Object Detection, and Critical View of Safety Assessment: Official Splits and Benchmark.* arXiv preprint arXiv:2312.12429, 2023.
42. Liu, R. and N. Choi, *A First Look at Wi-Fi 6 in Action: Throughput, Latency, Energy Efficiency, and Security.* Proceedings of the ACM on Measurement and Analysis of Computing Systems, 2023. **7**(1): p. 1-25.
43. Dabare, C., et al., *Differences in presentation, progression and rates of arthroplasty between hip and knee osteoarthritis: Observations from an osteoarthritis cohort study-a clear role for conservative management.* Int J Rheum Dis, 2017. **20**(10): p. 1350-1360.
44. Chai, C., et al., *Nutrient characteristics in the Yangtze River Estuary and the adjacent East China Sea before and after impoundment of the Three Gorges Dam.* Sci Total Environ, 2009. **407**(16): p. 4687-95.
45. Hakimi, A.A., et al., *A novel inexpensive design for high definition intraoperative videography.* Surgical Innovation, 2020. **27**(6): p. 699-701.
46. Nair, A.G., et al., *Surgeon point-of-view recording: using a high-definition head-mounted video camera in the operating room.* Indian journal of ophthalmology, 2015. **63**(10): p. 771-774.
47. Avery, M.C., *Intraoperative video production with a head-mounted consumer video camera.* Journal of Orthopaedic Trauma, 2017. **31**: p. S2-S3.
48. Wang, Z., et al., *Instance segmentation of point cloud captured by RGB-D sensor based on deep learning.* International Journal of Computer Integrated Manufacturing, 2021. **34**(9): p. 950-963.