Using Conditional Generative Adversarial Networks to Reduce the Effects of Latency in Robotic Telesurgery

Neil Sachdeva
Machine Perception and Cognitive Robotics Lab
Florida Atlantic University
Pine Crest School
Boca Raton, Florida, USA
neil.sachdeva@vulcan-ai.us

William Edward Hahn
Machine Perception and Cognitive Robotics Lab
Florida Atlantic University
Boca Raton, Florida, USA
williamedwardhahn@gmail.com

Misha Klopukh
Machine Perception and Cognitive Robotics Lab
Florida Atlantic University
Boca Raton, Florida, USA
mishakmak@gmail.com

Rachel St. Clair
Machine Perception and Cognitive Robotics Lab
Florida Atlantic University
Boca Raton, Florida, USA
rstclair2021@fau.edu

***Abstract*—The introduction of surgical robots brought about advancements in surgical procedures. The applications of remote telesurgery range from building medical clinics in underprivileged areas, to placing robots abroad in military hot-spots where accessibility and diversity of medical experience may be limited. Poor wireless connectivity may result in a prolonged delay, referred to as latency, between a surgeon's input and action a robot takes. In surgery, any micro-delay can injure a patient severely and in some cases, result in fatality. One way to increase safety is to mitigate the effects of latency using deep learning aided computer vision. While the current surgical robots use calibrated sensors to measure the position of the arms and tools, in this work we present a purely optical approach that provides a backup measurement of the tool position in relation to the patient's tissues. This research aimed to produce a neural network that allowed a robot to detect its own mechanical manipulator arms. A conditional generative adversarial network (cGAN) was trained on 1107 frames of a mock gastrointestinal robotic surgery from the 2015 EndoVis Instrument Challenge and corresponding hand-drawn labels for each frame. When run on new testing data, the network generated near-perfect labels of the input images which were visually consistent with the hand-drawn labels and was able to do this in 299 milliseconds. These accurately generated labels can then be used as simplified identifiers for the robot to track its own controlled tools. These results show potential for a conditional GANs as a reaction mechanism such that the robot can detect when its arms move outside the operating area in a patient. This system allows for more accurate monitoring of the position of surgical instruments in relation to the patient's tissue, increasing safety measures of a telesurgery system.**

***Index Terms*—Conditional Generative Adversarial Networks, Robotic Surgery, Latency**

I. Introduction

Surgical robots, such as the da Vinci Surgical System allow for surgeons to perform minimally invasive surgeries with pinpoint accuracy and complete maneuverability [1]. In a typical robotic surgery system, such as the da Vinci portrayed in figure 1, the surgeon's console is directly wired to the robot and a screen that shows a live feed of the robotic arms inside the patient.

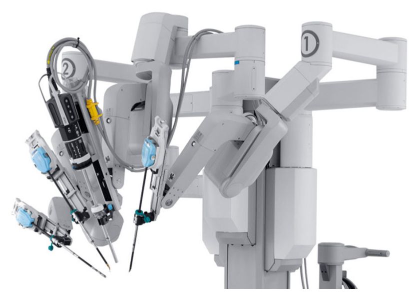


Fig. 1: Image of da Vinci Robot components

For surgical robots to have full reliability in a remote setup far from the operating surgeon, they need to be able to continue operating even in scenarios where network connection is unreliable, as any microsecond delay can potentially result in a serious accident. In addition, no networks have 100% reliability, so there is a lag time where a video feed could freeze or a command to move the robot is not received, in which case the robot would continue to keep moving even if the patient got in the way. These risks have discouraged the expansive use of the practice and while there are currently operating remote surgeries, they cannot be utilized on a large scale because of the potential dangers associated with latency [2]–[4]. Addressing the concern of latency is the primary concern of this work for aiding telesurgery reliability and practicality in the field.

By implementing a computer vision aided system to serve as an intermediary between the robot and the surgeon, the robot is no longer solely dependent on the surgeon and thus the effects of input lag are mitigated - specifically during the time it takes for a command to reach the robot which is when the on-board autonomous system can take control. In real-world applications, a robot would be stationed in a remote location and a doctor would be at their control station located in their own office. The neural network would be loaded onto the surgical robot's on-board computer and would be able to take control of the robot's arms whenever necessary. If a network interruption occurred, the neural network could recognize the robotic arms moving towards a dangerous position (i.e. colliding with the patient) and override the robot's controls and force it to stop. This system has the potential to accurately monitor the surgical instruments in relation to the patient's tissue. While the current surgical robots use calibrated sensors to measure the position of the arms and tools, in this work we present a purely optical approach backed by artificial neural networks that provides a backup measurement of the tool position in relation to the patient's tissues.

Convolutional Neural networks (CNN) are a subset of deep learning algorithms built to model the basic structure of the human primary visual cortex [5], [6]. In the case of image processing, they take images as inputs, learn features increasing in abstraction throughout the network layers, and then learn how these features relate to the specific image domain. This research focuses on a particular conditional generative adversarial network (cGAN) called Pix2Pix, as a potential use in robotic surgeries. Our novel contribution was to build a Pix2Pix GAN on novel data to recognize robotic arms in a surgical setting as a basis for an injury prevention system in a robotic telesurgery [7].

II. Theory

Conditional GANs combine a generative network that produces images from stochastic noise distributions, with a discriminator network that performs image recognition classification task [8]–[11]. These two networks compete against each other in a game-theory-like approach to see which network can become more accurate as shown in figure 2. The generator starts off by creating random noise images and feeding them into the discriminator, along with sample images from the original data set. The discriminator then decides whether the image it's fed is a 'real' picture - meaning from the data set - or a 'fake' image that was produced by the generator.

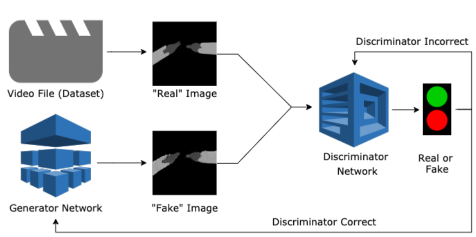


Fig. 2: High-level representation of cGAN algorithm

The generator G takes x and z then it produces y, with the goal of producing images that are indistinguishable from the training data. The discriminator D receives pairs (in and out) from both real images and fake images. The goal of the D is to distinguish between fake and real samples.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x|y)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{1}$$

The discriminator is a convolutional neural network architecture (CNN) that breaks down images in order to learn how to recognize and identify important parts of the image such as edges, corners, and colors through a process illustrated in figure 3. If it is fed an image from the generator, and it decides that the image is a fake, the generator takes that feedback and adjusts its weights in order to produce a more realistic image. The generator is a U-Net architecture that builds a dense embedding of an image using convolutional layers and expands that embedding into a new generated image [12]. This process is shown in figure 4. Eventually, the generator that started off producing random images, begins producing images that seem real enough to fool the discriminator. If the discriminator is wrong in its conclusion (for example guessing that a fake image was real), it will adjust its own weights to better its accuracy for the future.

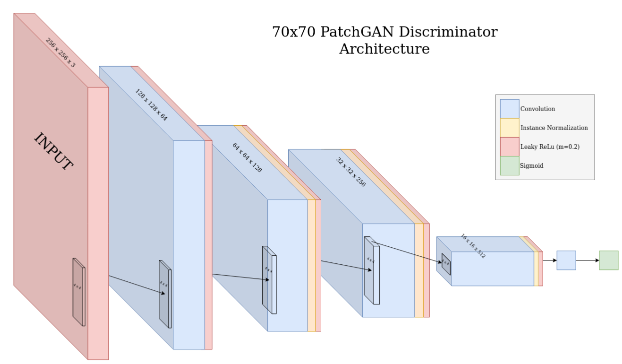


Fig. 3: Typical Convolutional Neural Network Architecture used for image processing, classification, and segmentation.

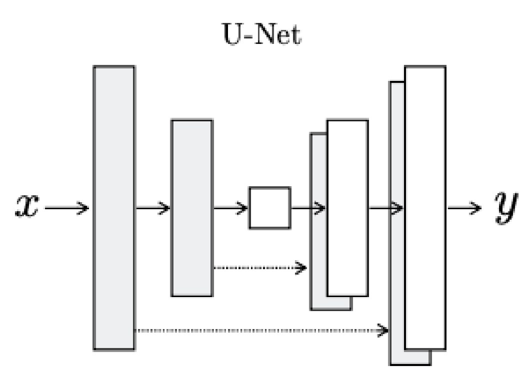


Fig. 4: U-Net encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

III. Data and Methods

A. Dataset

The 2015 Endovis Challenge dataset used as the training data of this work included videos of a mock gastrointestinal surgery in the form of 3 videos, each of which are 44 seconds long [13]. The first video is endoscopic video footage of a robotic arm moving around simulating a surgery in an ex-vivo setup. Each frame of the video has a corresponding hand-drawn label for the positioning of the right arm and left arm which makes up the other two video files (one video for the left arm segmentation, one for the right).

B. Data Preparation

This research utilized a PyTorch implemented Pix2Pix model written by Jun-Yan Zhu, Taesung Park, and Tongzhou Wang [7]. PyTorch is a Python-based deep learning framework modeled after the Torch framework that uses multidimensional arrays as tensors. Pix2Pix is a Conditional GAN that is specifically used for image to image translation and segmentation. It takes images and their labeled segmentations and learns how to convert from one to another. Because the entire research was conducted using Colab (Google's online Jupyter notebook), we were able to clone the Github repository to a Google drive and access the model from there.

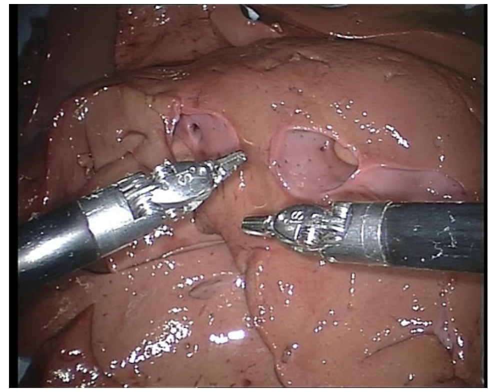
The model we used required the input data (images and labels) to be entered as singular paired images. We first split the video files into individual image frames pictured in figure 5a. Since the segmented label videos were separated by arm as shown in 5b and 5c, we combined them into one image with both arm segmentations show in 5d. The endoscopic pictures and labels were then stitched together to create an image with both frames side by side, which was then uploaded into the Google drive. This process was repeated 1107 times, for each frame of the video.

This model was trained for 200 epochs, and was tested for accuracy every 5 epochs.

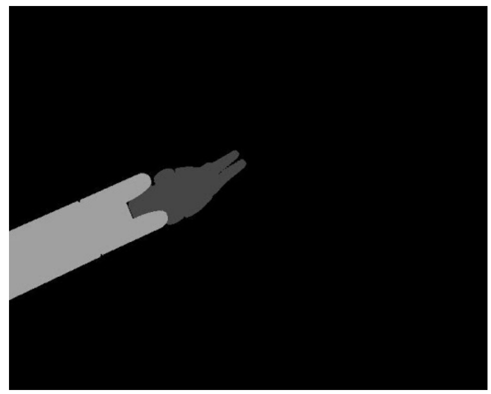
IV. Testing and Results

A. Generated Labels

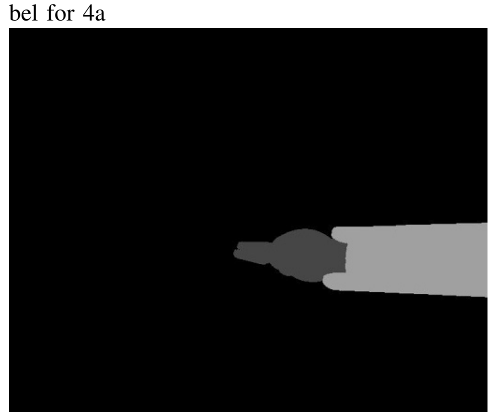
The main objective of this research was to produce labels for robotic surgery images. The primary results are the labels that were generated which are portrayed in table I. For the input image in the first epoch of training, the respective label produced by the generator was noticeably blurry and there is a stark difference when compared to its hand-drawn label for that same image shown. By the 200th epoch, the model got even more accurate and the label produced was nearly identical to the hand-drawn label.



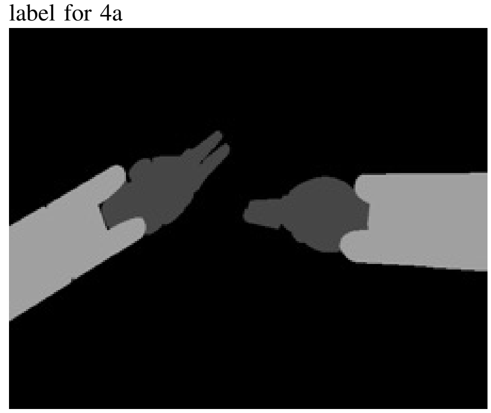
(a) Sample image of endoscopic video feed



(b) Left arm hand-drawn segmentation label for 4a



(c) Right arm hand-drawn segmentation label for 4a



(d) Combined segmentation label

Fig. 5: Sample frames from EndoVis Challenge dataset

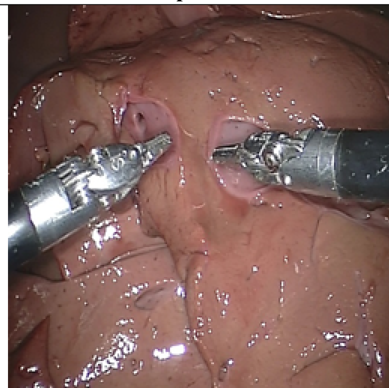
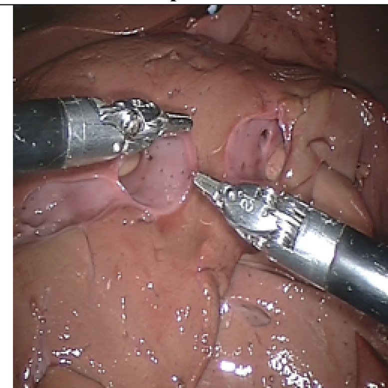
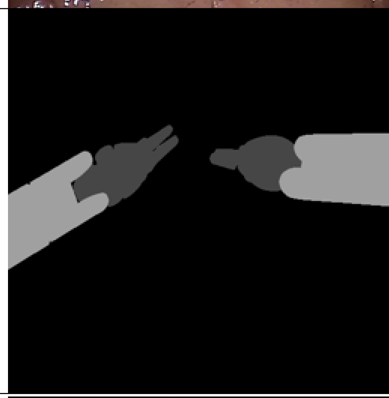
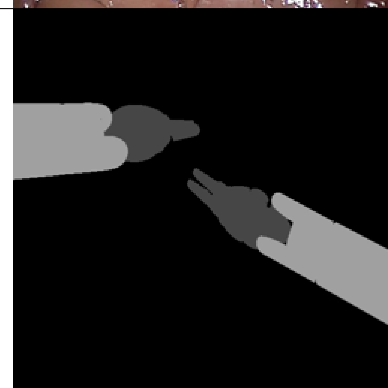
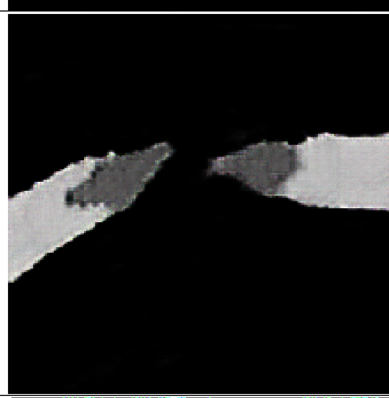
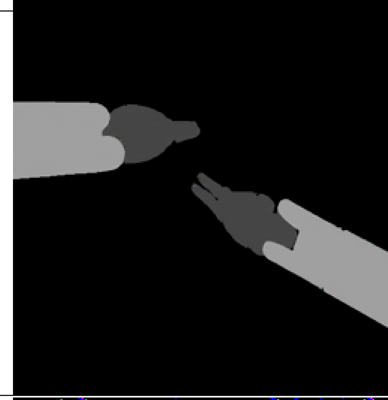
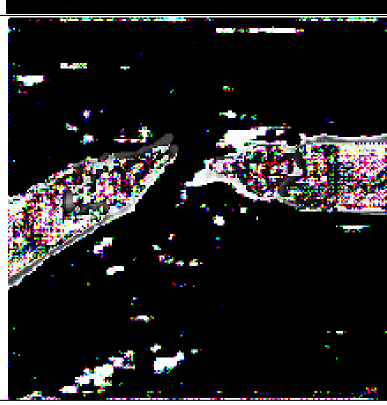
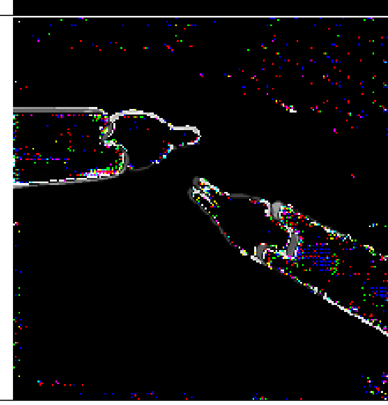
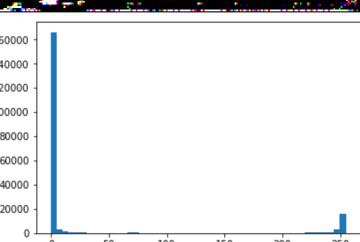
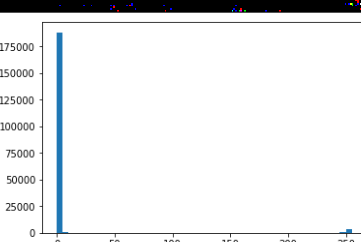
Description	Epoch 1	Epoch 200
Endoscopic camera frame		
Hand drawn label		
Label created by generator		
Subtracted difference between generated image and hand drawn image		
Histogram of per-pixel error		

TABLE I: Comparative results of untrained (epoch 1) vs trained (epoch 200) model

B. Image Subtraction

In addition to visually comparing the generated labels, we computed the mathematical differences between the trained model (epoch 200) and the untrained model (epoch 1). By subtracting the generated label from the hand-drawn label, we found the pixel differences between the two images. In table 1, the untrained model subtracted difference has a visibly large amount of noise in the image. While there is some noise on the background of the image (likely due to cropping) the majority of the pixel difference is centered in the robotic arms. This compared to the trained model in which the little noise that exists is mostly around the edges of the robotic arms. This data was then converted into a histogram in which we could see the pixel's color concentration within the subtracted images. For the untrained data, the spike of white pixels represented by the 250 values is noticeably higher than that of the trained data. This numerical representation of the data is conclusive evidence for the advancement of the model and its sophistication in the matter of recognizing robotic arms as we would expect the difference per pixel would decrease as images increased in similarity.

V. Discussion and Future Research

The network was able to perform very well on the surgical images with two arms, achieving near-perfect accuracy with the generator by epoch 200. In figure 6, the difference in the number of non-zero pixels is highlighted, showing a five-fold increase in accuracy. This supports the hypothesis that a Conditional Generative Adversarial Network has the capability to learn and reproduce what a surgical robotic arm looks like in a surgical setting.

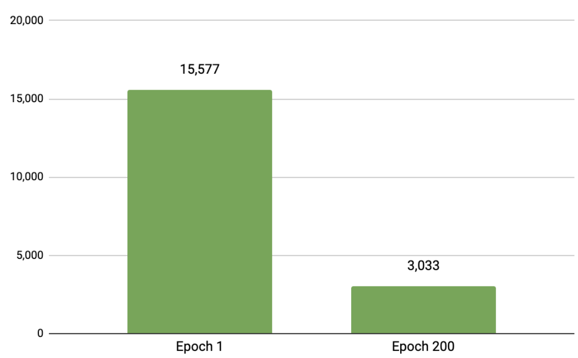


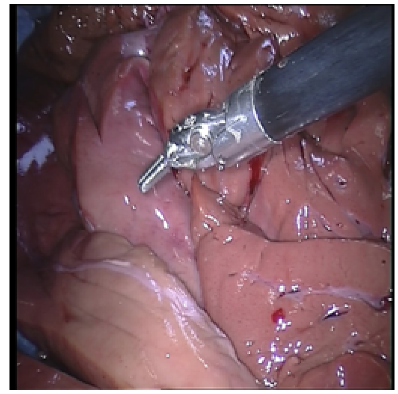
Fig. 6: Number of nonzero pixels in the subtracted difference images (comparing epochs 1 to 200)

With the ability to segment and track the robotic arms, the next important piece of this research is the time factor. If the trained model took too long to process the images that it was given, then the entire premise of using it as a solution to the issue of latency in remote surgery would fail. To test this, we wrote a script that timed how long the model took to segment a single input image which ended up being 299 milliseconds.

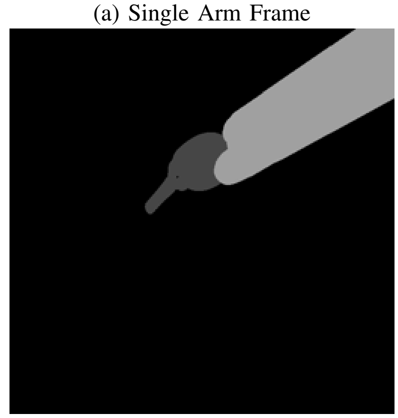
The findings presented in this research indicate that a neural network of the conditional generative adversarial architecture can be effectively used to teach a system how to recognize its own robotics limbs; however, there are some limitations of the model that need to be addressed before this system can be applied to a real surgery.

The dataset that we used for our model training was limited to images of two robotic arms moving around in gastrointestinal surgery. Because of this, the model learned that there were always going to be two arms in every image, and when it was tested on an image with only one robotic arm, the generator got confused and produced an inaccurate image.

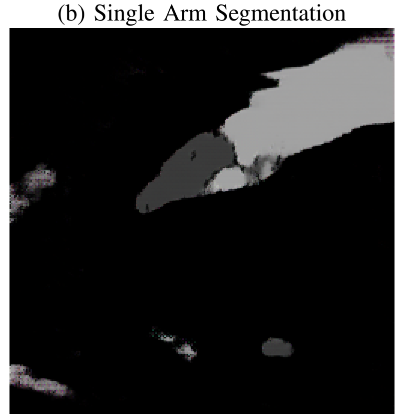
The versatility that deep learning allows makes it possible to expand the training data to include images of a single-



(a) Single Arm Frame



(b) Single Arm Segmentation



(c) Single Arm Generated Segmentation

Fig. 7: Test frames to check for adaptability of model

armed robot, and the model will learn how to recognize them. In fact, many issues can simply be solved by adding to the training data and familiarizing the model with all types of robotic surgeries. For example, if a different attachment such as a stapler or forcep needs to be added, the model can be trained to recognize all of the necessary components by adding pictures of them to the database.

This network has the potential to allow the application of telesurgery in places where high-speed fiber-optic connections are not available, such as underdeveloped countries, on a submarine, or in outer space. Telesurgery can be applied to the case where a wounded soldier needs a surgery that requires they are usually flown to the closest hospital, but for many war zones, these doctors are hard to find or too far away for the injured to reach. Since the start of the Afghanistan war in 2001, there have been over 2,300 American casualties, and more than 20,000 wounded from service, many of whom might have been saved if proper access to medical facilities were available [14]. A telesurgery system such as the one pictured in figure 8 will give medical professionals further reach to help patients, and can allow telesurgery to save lives in the years to come.

The aim of this research was to produce a system that could learn how to recognize robotic limbs, however, the potential of cGANs in surgery has a much larger scope that has yet to be explored. Applications in organ labeling to improve accuracy and tracking of other surgical instruments can all be achieved with neural networks and machine learning. In this research, we were able to produce a neural network that has the ability to track robotic arms and tell their position in a surgical context. By devising a system to detect when the arms are moving towards dangerous positions within patients, this research will provide the base for future research in applying telesurgery to places where high-speed fiber-optic connections are not available.

The network will continue to be trained on different types of robotic surgery data and pre-trained models will also be available open source for further research. All code can be found at https://github.com/Neil-Sachdeva/RoboGAN.

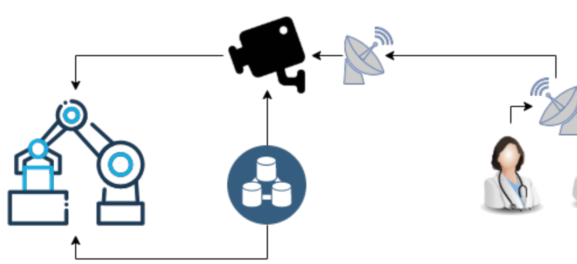


Fig. 8: Model setup for Robotic Telesurgery - cGAN serves as an interface between the doctor receiving live video data from the camera and the robot performing surgery

References

[1] "Intuitive surgical davinci surgical robot."

[2] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, *et al.*, "Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge," *IEEE transactions on medical imaging*, vol. 36, no. 6, pp. 1231–1249, 2017.

[3] Y. Jin, *Towards Intelligent Surgery: Dynamic Surgical Video Analysis with Deep Learning*. PhD thesis, The Chinese University of Hong Kong (Hong Kong), 2019.

[4] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, pp. 2278–2324, 1998.

[6] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, p. 101552, 2019.

[9] S. Kazeminia, C. Baur, A. Kuijper, B. van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "Gans for medical image analysis," *arXiv preprint arXiv:1809.06222*, 2018.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[12] S. Kamrul Hasan and C. A. Linte, "U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instrument," *arXiv preprint arXiv:1902.08994*, 2019.

[13] "Endovissub-instrument - grand challenge."

[14] "Coalition casualties in afghanistan," Nov 2019.

[15] M. Pfeiffer, C. Riediger, J. Weitz, and S. Speidel, "Learning soft tissue behavior of organs for surgical navigation with convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1147–1155, 2019.

Appendix

Additional results from model tests

Camera Image | Real Label | Generated Label