Digital Twin of a Multi-Arm Robot Platform based on Isaac Sim for Synthetic Data Generation

Juan José Quiroz-Omaña, Murilo Marques Marinho, Member, IEEE, and Kanako Harada, Member, IEEE

***Abstract*— Autonomous robotic surgery combines state-of-the-art strategies to potentially provide more efficacy and safety regardless of the surgeon's skill. These approaches usually use CNNs, which require a large amount of data for suitable training. However, in some applications as medical procedures on animals, performing thousands of trials would be ethically hard to justify. In this letter, we develop a digital twin on Isaac Sim to create a synthetic dataset. We use synthetic images to train a CNN for image segmentation tasks of our AI robot science (AISP) platform. We compare it to a second CNN trained with real images showing that for a specific validation dataset, the CNN trained with synthetic images performs similarly to the one trained with real images.**

The dataset and trained models are freely available for noncommercial use at `https://github.com/AISciencePlatform/icra2023_synthetic_data_pretraining_for_robotics`.

I. Introduction

Surgical robots are platforms designed to perform challenging medical procedures in highly constrained workspaces [1]. Usually, surgeons use them as smart tools [2], in which the robots are under their teleoperator control. This strategy exploits the robot's dexterity and reduces the surgeon's hand tremor providing more efficacy and safety. In this paradigm, the quality and accuracy of robot-assisted surgery (RAS) is related to the surgeon's skills, which are subjected to human imprecision. However, new developments in artificial intelligence, computer vision, and motion control techniques have enabled greater robot autonomy [3].

Recently, we are developing an AI-robot Science Platform (AISP) [4], which is a multi-arm robotic platform composed of four serial manipulatorsdesigned for scientific experimentation. The goal we envision is to provide different levels of autonomy to perform tasks alongside humans. Different from teleoperated RAS, autonomous robotic surgery (ARS) can potentially provide efficacy and safety regardless of the surgeon's skill [3] at expense of removing the human operator, which usually can use their vision to compensate for kinematic inaccuracies [5]. Because of that, ARS approaches could require additional strategies to deal with it. The first step toward this direction is to improve the accuracy of the platform by means of adaptive control strategies taking advantage of DNN semantic segmentation approaches.

Motion control has been addressed in strategies that use the robot model to describe its configuration in the workspace. The model usually encapsulates information about the robot's geometric parameters, the robot's inertial parameters, relative poses between different reference frames, and so on. Consequently, the overall accuracy is directly related to the model quality, and therefore, in applications that demand higher levels of accuracy as assistive surgical robotics, calibration procedures could be required.

Marinho et al. [6] proposed an adaptive kinematic control strategy based on quadratic programming. This strategy enables the use of any sensor that provides partial[1] or complete[2] task-space measurements to compensate online for calibration errors. Such calibration strategy can use the information from the image using instrument/detection tracking strategies, in which semantic segmentation is usually an important component [5], [7], [8].

In such data-driven approaches based on convolutional neural networks (CNNs), it is well-known that they are powerful but require a large amount of data for suitable training [9] and annotated data is often the bottleneck. Furthermore, in our target application of medical procedures on animals, performing thousands of trials would be inconceivable in terms of logistics, cost, and ethics. Also in other fields with similar challenges, the use of synthetic datasets is trending [10]. Simulators enable the creation of realistic images and perfectly annotated data using state-of-the-art rendering techniques. Furthermore, synthetic image generation allows a large degree of domain randomization [11], which is paramount to closing the gap between synthetic and real data [10]. In our previous work [12], we have shown that in robot-aided endonasal surgery, increased rendering realism positively correlates with the quality of the output in real images, using networks trained only on synthetic data.

In this work, further develop our IsaacSim[3]-based digital twin of AISP to autonomously generate synthetic datasets. We use the generated datasets to train a CNN to perform image segmentation tasks and evaluate the results against another CNN trained on real images. Our goal is to obtain information about the robots to estimate their configuration [13], [14]. Nonetheless, such estimation is outside the scope of this paper.

This project was funded by JST Moonshot R&D (JPMJMS2033-09).

M. M. Marinho is with the University of Tokyo, Department of Mechanical Engineering, Tokyo, Japan. e-mails: murilo@g.ecc.u-tokyo.ac.jp.

J. J. Q. Omaña and K. Harada are with the University of Tokyo, Graduate Schools of Medicine & Engineering, Tokyo, Japan. e-mails: {juanjqo,kanakoharada}@g.ecc.u-tokyo.ac.jp.

[1] E.g., position, orientation, and distance.

[2] Pose, that is, combined position and orientation.

[3] https://developer.nvidia.com/isaac-sim

II. Methodology

The AISP platform is composed of two collaborative-type Cobotta arms (CVR038, Densowave, Japan), and two industrial manipulators (VS050, Densowave, Japan). Each robot is equipped with a unique customized end-effector aiming to enable a large set of applications. One cobotta has a customized micro-drill, whereas the other one has a grasper for handling regular-sized cotton swabs. The industrial manipulators are equipped with customized actuators based on a rotary gripper module to operate tweezers and scissors.

The goal is to evaluate a CNN trained with synthetic images for segmentation tasks of the AISP's Cobotta arms. In order to compare its performance, we trained a second CNN with real images. Both CNNs are based on U-NET [15] and are labeled as *UNET-R* and *UNET-S.* The former denotes the one trained with real images, whereas the latter denotes the one trained with synthetic images. Fig. 1 shows the overview of the proposed approach.

A. Synthetic dataset

To generate the synthetic dataset of our robotic platform we created a digital twin based on Isaac Sim.[4] We used the platform's CAD models and defined suitable render materials to take into account the specular reflection of the metallic surfaces, the acrylic panels on the doors, and the robots. Furthermore, we replicated the light conditions as in the real environment. Figure 2 shows the real platform and its digital twin side-by-side. Using the simulator, we generated 10000 greyscale images[5] of 512 x 512 pixels with their respective segmented images. We used RTX – Interactive (Path Tracing) render and Isaac Sim Replicator tools to uniformly randomize the pose of the camera and the joint positions of both Cobotta manipulators,[6] as shown in Fig. 3. We fixed a start pose of the camera to match the framing of the real images and performed variations in the *x*, *y,* and *z* axes of 10cm, 0.2cm, and 4cm respectively. In all cases, the camera orientation was pointed to the table. The rendering of all images took about 14 hours in a computer running Windows 11 64 bits, equipped with a Intel i9-13900K with 64GB RAM, and two RTX-A6000 Ada Generation.

B. Real dataset

To generated the real dataset we performed a teleoperation drilling task[7] and captured 100 snapshots using a 4K camera (STC-HD853HDMI, Omron-Sentech, Japan). All images were converted to greyscale images of 512 x 512 pixels, as shown in Fig. 3. We manually segmented each image. We used 20 images as test dataset. Therefore, the real training dataset comprises 80 images.

III. Evaluation

To validate the proposed approach, we compared the predictions of both CNNs with respect to the test dataset, as shown in Fig. 1. We use Pytorch-UNet[8] on Python 3.10 with PyTorch 1.13.1+cu117 to train both CNNs.

Figure 4 shows the box-and-whiskers plot of the IoU $\in [0,1]$ and the Dice $\in [0,1]$ similarity coefficients for both CNNs. A similarity of 1 represents a perfect match between the predicted image and the real one. In both cases the predicted images are accurated. However, the performance of *UNET-R* is higher than *UNET-S.* This is expected since our synthetic dataset does not replicate perfectly all details contained in the real system.

Figure 5 shows the qualitative results of the image segmentation at the best and worst cases for *UNET-S.* Some shaded regions of the robots that resemble the table, and very bright regions of pixels, such as lamplight, are features of the real dataset that were not represented in the synthetic one. Consequently, those regions were very challenging to *UNET-S,* as expected.

IV. Conclusions

In this preliminary work, we evaluated the accuracy of a CNN trained with synthetic images to perform image segmentation tasks of our robotic platform. The synthetic dataset comprises 10000 rendered images with random joint positions and random camera poses of our digital twin, which we created on Isaac Sim. Furthermore, we trained a second CNN using 80 real images and compared both with respect to our ground truth composed of 20 different real images. The results showed that for this small test, the CNN trained with synthetic images performed similarly to the one trained with real images despite the missing details of the real system in the synthetic dataset.

Future works will be focused on an ablation study to determine the relevant parameters to be randomized in the generation of the synthetic dataset. Furthermore, we want to explore strategies for recognition of occluded objects.

References

[1] P. E. Dupont, B. J. Nelson, M. Goldfarb, B. Hannaford, A. Menciassi, M. K. O'Malley, N. Simaan, P. Valdastri, and G.-Z. Yang, "A decade retrospective of medical robotics research from 2010 to 2020," *Science Robotics*, vol. 6, no. 60, Nov. 2021. [Online]. Available: https://www.science.org/doi/10.1126/scirobotics.abi8017

[2] R. Taylor, "A Perspective on Medical Robotics," *Proceedings of the IEEE*, vol. 94, no. 9, pp. 1652–1664, Sep. 2006. [Online]. Available: http://ieeexplore.ieee.org/document/1717783/

[3] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Leonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science Robotics*, vol. 7, no. 62, Jan. 2022. [Online]. Available: https://www.science.org/doi/10.1126/scirobotics.abj2908

[4] M. Marques Marinho, J. J. Quiroz-Omaña, and K. Harada, "Design and validation of a multi-arm robotic platform for scientific exploration," 2022. [Online]. Available: https://arxiv.org/abs/2210.11877

[4] We used Isaac Sim 2022-2-1.

[5] We used 8k images for training and 2k for validation.

[6] We set suitable random camera poses and robot configurations to obtain similar rendered images to the real dataset.

[7] For this task, we used the Cobotta manipulators only.

[8] https://github.com/milesial/Pytorch-UNet

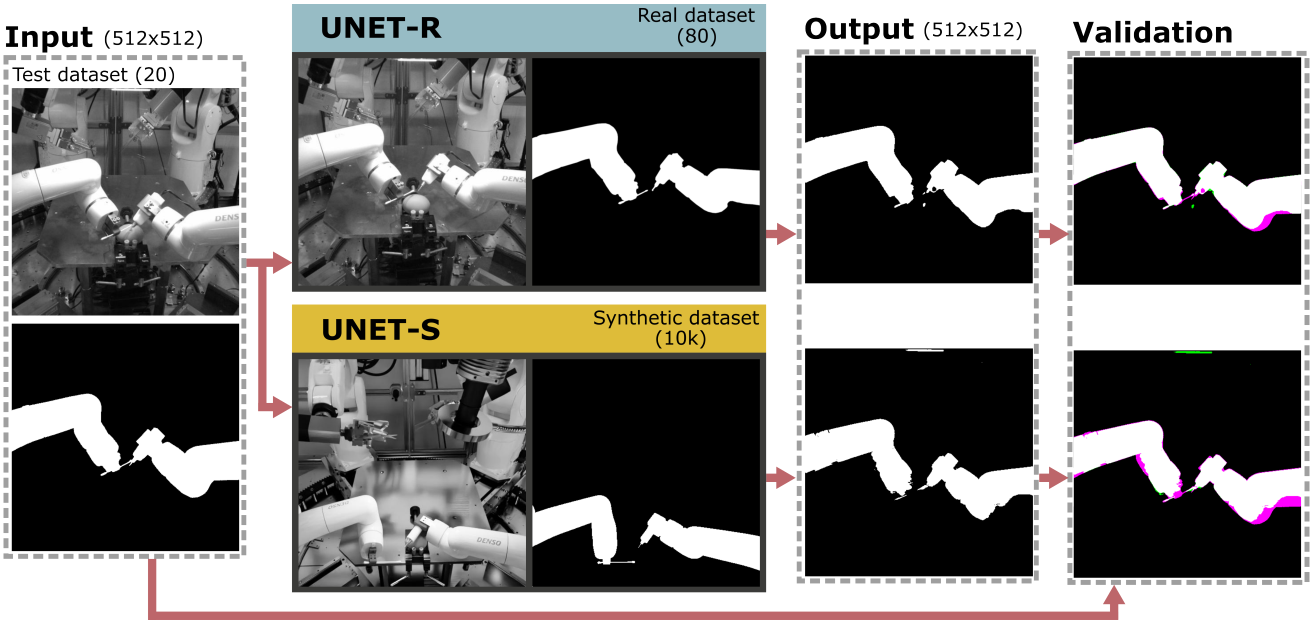

Fig. 1. Overview of the proposed strategy. The predictions of a CNN trained with 80 real images (*UNET-R*) and a CNN trained with 10000 synthetic images (*UNET-S*) are compared with a test dataset.

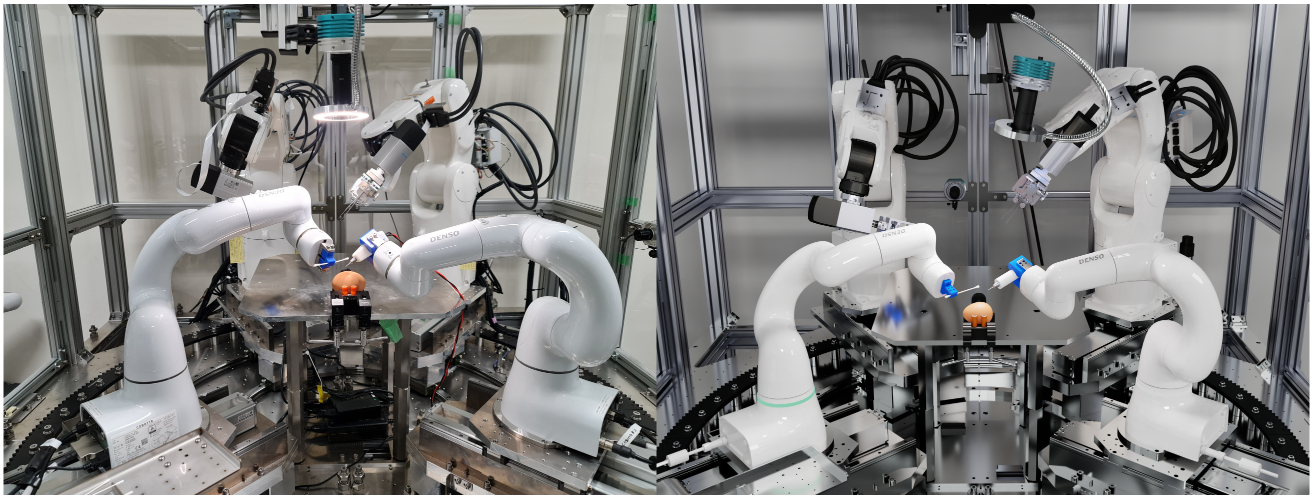

Fig. 2. On the *right*, our robotic platform. On the *left*, a snapshot of the Isaac Sim scene using path tracing render.

[5] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi, "Single-Shot Pose Estimation of Surgical Robot Instruments' Shafts from Monocular Endoscopic Images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, May 2020, pp. 9960–9966. [Online]. Available: https://ieeexplore.ieee.org/document/9196779/

[6] M. M. Marinho and B. V. Adorno, "Adaptive Constrained Kinematic Control Using Partial or Complete Task-Space Measurements," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3498–3513, Dec. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9805834/

[7] Y. Sun, B. Pan, and Y. Fu, "Lightweight Deep Neural Network for Real-Time Instrument Semantic Segmentation in Robot Assisted Minimally Invasive Surgery," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3870–3877, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9380950/

[8] Z.-L. Ni, G.-B. Bian, Z.-G. Hou, X.-H. Zhou, X.-L. Xie, and Z. Li, "Attention-Guided Lightweight Network for Real-Time Segmentation of Robotic Surgical Instruments," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. Paris, France: IEEE, May 2020, pp. 9939–9945. [Online]. Available: https://ieeexplore.ieee.org/document/9197425/

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015. [Online]. Available: http://www.nature.com/articles/nature14539

[10] C. M. de Melo, A. Torralba, L. Guibas, J. DiCarlo, R. Chellappa, and J. Hodgins, "Next-generation deep learning based on simulators and synthetic data," *Trends in Cognitive Sciences*, vol. 26, no. 2, pp. 174–187, Feb. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S136466132100293X

[11] M. Yoshimura, M. M. Marinho, K. Harada, and M. Mitsuishi, "MBAPose: Mask and Bounding-Box Aware Pose Estimation of Surgical Instruments with Photorealistic Domain Randomization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Prague, Czech Republic: IEEE, Sep. 2021, pp. 9445–9452. [Online]. Available: https://ieeexplore.ieee.org/document/9636404/

[12] S. A. Heredia Perez, M. Marques Marinho, K. Harada, and M. Mitsuishi, "The effects of different levels of realism on the training of CNNs with only synthetic images for the semantic segmentation of robotic instruments in a head phantom,"

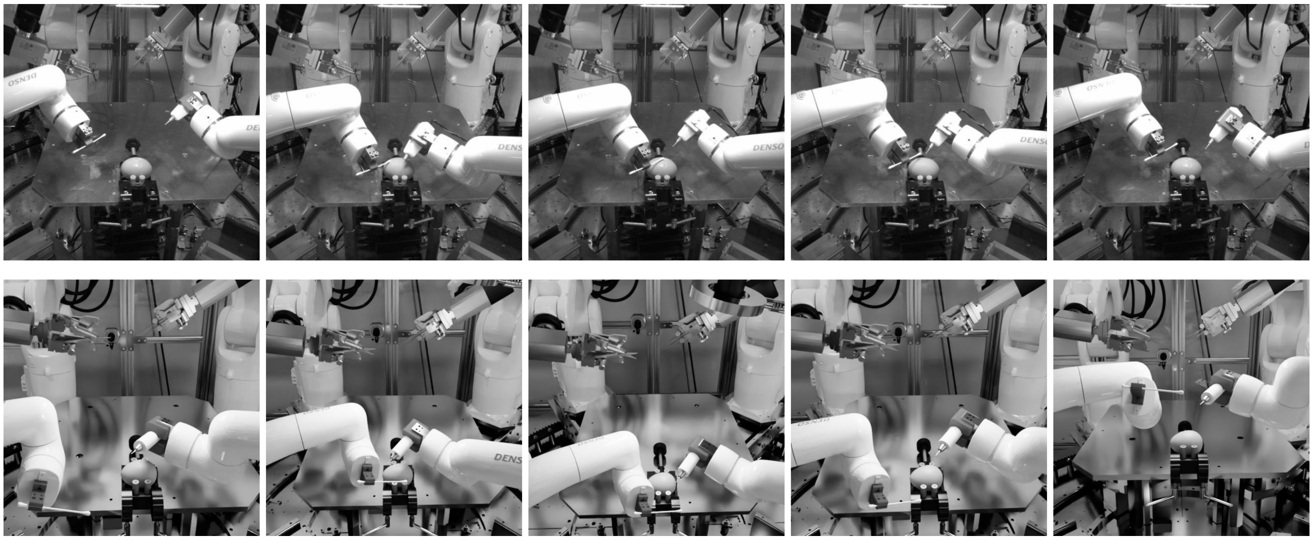

Fig. 3. Dataset samples. On the *top*, the real dataset, which is composed of snapshots of a teleoperated drilling task using both Cobotta arms. On the bottom, the synthetic dataset. In this case, both the camera pose and the configuration of both arms were randomized.

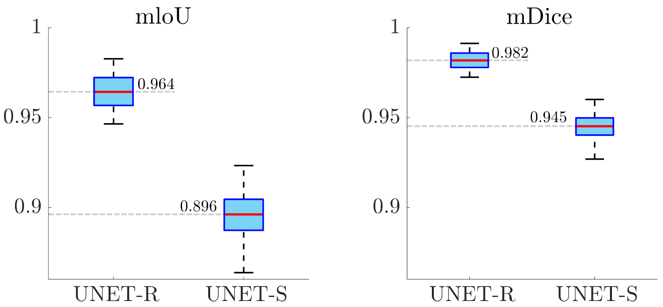

Fig. 4. Similarity coefficients of the network trained with real images (*UNET-R*) and the one trained with synthetic images (*UNET-S*). On the *left,* the box-and-whiskers plot of the intersection over union (IoU). On the *right*, the box-and-whiskers plot of the Dice coefficient. In this pilot study, *UNET-S* showed good enough performance despite being trained with only synthetic images.

International Journal of Computer Assisted Radiology and Surgery, vol. 15, no. 8, pp. 1257–1265, Aug. 2020. [Online]. Available: https://link.springer.com/10.1007/s11548-020-02185-0

[13] F. Widmaier, D. Kappler, S. Schaal, and J. Bohg, “Robot arm pose estimation by pixel-wise regression of joint angles,” in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. Stockholm: IEEE, May 2016, pp. 616–623. [Online]. Available: http://ieeexplore.ieee.org/document/7487185/

[14] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, “Single-view robot pose and joint angle estimation via render & compare,” 2021.

[15] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” May 2015, arXiv:1505.04597 [cs]. [Online]. Available: http://arxiv.org/abs/1505.04597

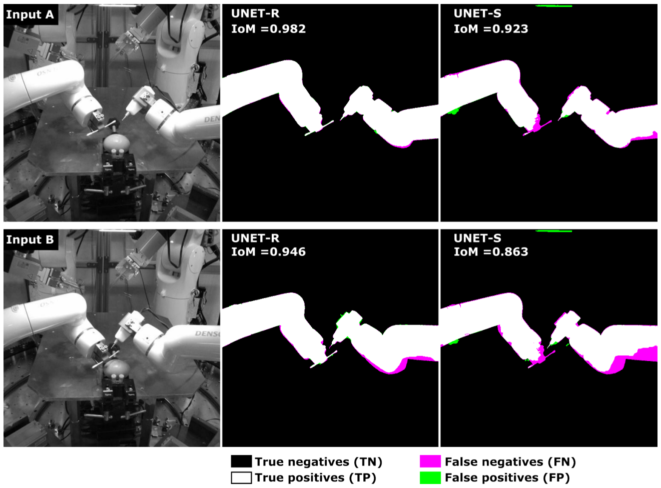

Fig. 5. Representative results of the semantic segmentation for two samples. On the *top*, the results where *UNET-S* (Synthetic case) performed better (IoU = 0.923). On the *bottom*, the results where *UNET-S* had the lowest score (IoU = 0.863). The white region represents the correct segmentation of both Cobotta manipulators (TP) whereas the black region represents the correct segmentation of the pixels outside of them (TN). Furthermore, the green color represents regions wrongly considered as part of the Cobottas (FP), whereas the magenta color denotes regions wrongly considered outside of them (FN).