Image Synthesis with Generative Adversarial Networks to Augment Tool Detection in Microsurgery

Mastaneh Torkamani-Azar[1] MASTANEH.TORKAMANI@UEF.FI
YuChun Liu[*1] YUCHLIU@STUDENT.UEF.FI
Jani Koskinen[†1] JANI.KOSKINEN@UEF.FI
Ahmed Hussein[2,3] AHMED.HUSSEIN@KUH.FI
Matti Iso-Mustajärvi[2,4] MATTI.ISO-MUSTAJARVI@KUH.FI
Hana Vrzakova[1] HANA.VRZAKOVA@UEF.FI
Roman Bednarik[1] ROMAN.BEDNARIK@UEF.FI
[1] *School of Computing, University of Eastern Finland, Joensuu, Finland.*
[2] *Microsurgery Center, Kuopio University Hospital, Kuopio, Finland.*
[3] *Department of Neurosurgery, Faculty of Medicine, Assiut University, Assiut, Egypt.*
[4] *Otorhinolaryngology Outpatient Clinic, Kuopio University Hospital, Kuopio, Finland.*

Editors: Under Review for MIDL 2022

Abstract

For the first time in literature, we investigate the capability of Generative Adversarial Networks (GAN) for synthesizing realistic images of microsurgical procedures and augmenting training data for surgical tool detection. We employ videos from practice and intraoperative neurosurgical procedures to train and evaluate two recent GAN models that have shown promise in high-resolution image generation: StyleGAN2 with Adaptive Discriminator Augmentation and StyleGAN2 with Differential Augmentation. Models were trained with limited data for both conditional and unconditional image generation, where the conditional models generated images with and without surgical tools. Our results show that the unconditional models achieved FID scores between 6 and 25 units lower than the conditional models for the two practice datasets. The best performance (FID = 42.16 and 25.17) was achieved in the Go-around practice task and was comparable to the previous benchmark performance of StyleGAN2 with Differential Augmentation. Experts' visual inspection showed that while synthetic images had faults that exposed their true origin to the human eye, a sizable portion of them included identifiable surgical instruments. Experiments with object detection showed that augmenting the training data with synthetic microsurgical data improved the mean average precision for detecting tool tips in practice microsurgery datasets by 3%. Future work will include improving the quality of image synthesis and investigating key visual cues in expert assessment of surgical scenes for applications in robust surgical tool detection, bimanual skill evaluation, and surgical phase understanding in microsurgery.

Keywords: Surgical tool detection, AI-assisted surgery, microsurgery, generative adversarial networks, StyleGAN, data augmentation, limited data, nearest neighbors.

1. Introduction

High-fidelity surgical scene imagery and their abundance are important for developing intelligent medical systems. Clinical applications, such as surgical process understanding, image-guided navigation, instrument tracking, and telematic surgery demand high-quality images

[*] Contributed equally
[†] Contributed equally

(Chadebecq et al., 2020; Kennedy-Metz et al., 2021). Computer vision can drive novel diagnostics, uncover a wealth of information during surgical procedures, and power innovative intraoperative solutions for surgical operation rooms (OR). The successful computer-vision applications in the OR relies on robust, automated tool detection from surgical recordings (Philipp et al., 2021). Similar to other medical-imaging applications, surgical-tool detectors are highly sensitive to the quality and quantity of training data (Torres-Velazquez et al., 2021). However, even when privacy concerns are resolved, curating reliable surgical datasets is time consuming and computationally inefficient. These challenges hinder the access to representative data for developing scalable and intelligent surgical systems (Maier-Hein et al., 2022).

Automated detection of surgical tools is a novel technique in computer-assisted surgery that enables action recognition and objective assessment of surgical expertise (Belykh et al., 2018; Philipp et al., 2021; Davids et al., 2021; Koskinen et al., 2022). Experimental studies with eye trackers have shown that tool tips are important regions of interest in neurosurgery due to their role in distinguishing the level of surgical expertise (Eivazi et al., 2012, 2017). However, operation fields are often highly magnified –as is the case in microsurgery– and suffer from blurriness, uneven illumination, and tool tip occlusion by tissues and other objects (Leppänen et al., 2018; Yamazaki et al., 2020; Shi et al., 2020), making tool detection a significant challenge. In neurosurgery and ophthalmic surgery where a standard approach is to operate in narrow microsurgical openings with microinstruments, low and non-uniform illumination, coaxial instrument positions, and excessive instrument movements hinder robust video-based detection of tool tips (Leppänen et al., 2018).

In daily practices, surgeons record and edit surgical videos to capture the key surgical phases. However, video editing inadvertently reduces the amount of data available for training automated instrument detection models (Vedula and Hager, 2017; Koskinen et al., 2022). Furthermore, the majority of surgical-tool detectors are built upon datasets with inadequate diversity in tools and settings (Davids et al., 2021). To overcome the limited size and diversity in microsurgical image datasets, generative adversarial networks (GANs) have shown remarkable data augmentation capability in applications such as brain-tumor segmentation (Calimeri et al., 2017; Shin et al., 2018; Quiros et al., 2020).

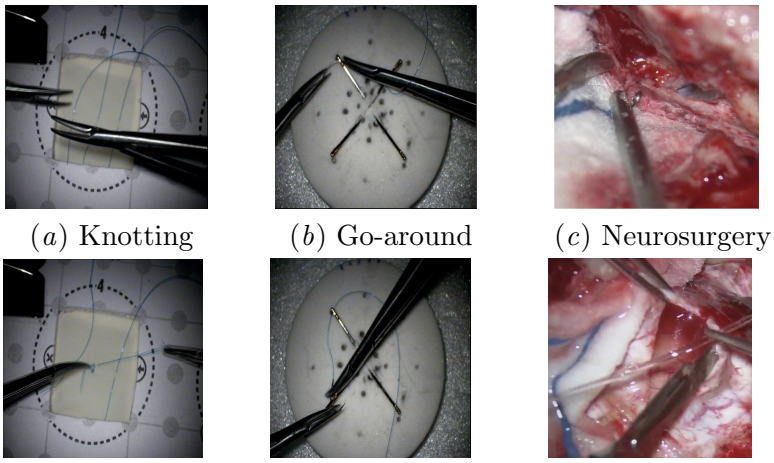
Contributions. We are the first to report on the potentials of GANs in microsurgery. Two variants of the StyleGAN2 are utilized to synthesize images of microsurgical instruments and brain tissues. The images are generated using unconditional and conditional GANs. In unconditional GANs, the task is to synthesize realistic images, whereas in conditional GANs the synthesized images are also automatically labeled("with surgical tools" and "without surgical tools"). All models are trained on videos from microsurgical practice and neurosurgery under authentic OR settings (i.e., large variation in illumination, contrast, color, and instruments). The quality of synthetic images and generated tools is assessed using Fréchet Inception Distance and Kernel Inception Distance, and by two microsurgeons manually. To demonstrate that GANs have refrained from simply memorizing input images, a nearest neighbor-based algorithm compares synthetic images to real training datasets in the pixel space. Finally, we conduct two classification experiments to demonstrate improvement in tool-tip detection using data augmentation. To our best knowledge, this is the first attempt to apply GANs in microsurgery and to assess the potentials of conditional and unconditional image generation in separating the tools from background scenes.

2. Materials and Methods

2.1. Source Datasets

Two source video datasets from microsurgery were used for image synthesis (Table 1). The *Practice Dataset* was collected during microsurgery training where participants completed various bimanual tasks using microforceps and a needle holder: surgical knotting (or suture tying), go-around (or bimanual handling), and object alignment under a surgical microscope (Zeiss Omni Pico, 15 FPS). These tasks represented fundamental components of surgical skills (Siu et al., 2010). In this work, we used videos from the Knotting and Go-around tasks of two participants and labeled a total of 3141 images with tools and 310 images without tools in Knotting, and 3124 images with tools and 235 images without tools in the Go-around task. The original image size was 720×486 px was downsampled to 512×512 px (denoted as Practice-512-Train dataset). Next, redundant and unfocused images were discarded that resulted in a dataset of 655 images with tools and 150 images without tools in the Knotting task, and 635 images with tools and 170 images without tools in the Go-around task. These images were downsampled to 256×256 px (denoted as Practice-256-Train dataset).

The *Intraoperative Dataset* (INT) comprised one video of an authentic neurosurgery with the anterior transpetrosal approach (Morisako et al., 2019). The INT Dataset contained complex backgrounds, higher tool variability, and fewer images due to being heavily edited at the source. In addition, the video included microscope movements and rapid changes in illumination, color, contrast, motion blur, and partial blockage by surgeons' hands. After visual inspection and elimination of low-quality, redundant, and highly blurry images, the amount of available data became limited. Random cropping with $n = 3$ was applied on the original images of 1280×720 px for data augmentation. The images were downsampled to 512 px and 256 px resolutions. Figure 1 demonstrates examples for each task.



(*a*) Knotting (*b*) Go-around (*c*) Neurosurgery

(*d*) Knotting (*e*) Go-around (*f*) Neurosurgery

Figure 1: Source images from Practice and Intraoperative microsurgical datasets. The top row illustrates high-quality images while the bottom row displays various image artifacts: (d) blurriness and out-of-focus, (e) complex poses, close tool tips, and low needle visibility, and (f) multiple tools in complex anatomical backgrounds.

2.2. Data Augmentation using StyleGAN2 Extensions

StyleGAN and StyleGAN2 represent state-of-the-art architectures in high-resolution image synthesis (Karras et al., 2019, 2020b). StyleGAN2 relies on large computational resources and training datasets (10^5-10^6 images). In small training datasets, StyleGAN2 adapts too quickly which can lead to "Mode Collapse" (Goodfellow et al., 2020; Arjovsky et al., 2017). In addition, data augmentation (i.e., cropping, flipping, scaling, and color transformation) used by StyleGAN2 can leak into synthesized images. To overcome these shortcomings, StyleGAN2+Ada adaptively stabilizes during training (Karras et al., 2020a). StyleGAN2+DiffAugment augments both real and generated samples used in the discriminator and propagates the gradients of augmented samples to the generator. Both models have demonstrated their success by achieving FID scores 2-4 times lower than the original StyleGAN2 once trained with 1000 samples (Zhao et al., 2020).

Table 1: Characteristics of microsurgery training datasets. Images in Knotting, Go-around, and authentic neurosurgery were manually labeled and used in image synthesis.

Dataset	Task	Tools [n]	No Tools [n]	Total [n]	Synthesized
Practice-256-Train	Knotting	655	150	805	200
Practice-256-Train	Go-around	635	170	805	200
INT-256-Train	Neurosurgery	237	-	237	50
Practice-512-Train	Knotting	3141	310	3451	301
Practice-512-Train	Go-around	3124	325	3449	301
INT-512-Train	Neurosurgery	232	-	232	301

2.3. Problem Formulation: Conditional and Unconditional Image Synthesis

GANs are unsupervised models while conditional GANs (CGAN) are supervised models that can output data with given class labels (Mirza and Osindero, 2014). This additional information enables the networks to learn the simplified, class-conditional distributions instead of the overall distribution of entire training sets (Kaneko et al., 2017). In this work, experiments are run with both unconditional and conditional models to test if class information helps training GANs with small datasets. Selecting two classes for images with and without tools is motivated by a previous work where the model had limited exposure to "empty" images (Koskinen et al., 2022), and is expected to tackle the imbalanced class distribution.

2.4. Evaluation Metrics

Three quantitative (FID, KID, and average Manhattan distance) and one qualitative metrics were used to compare the synthesized images to real images. FID measures the difference between synthetized and real images as a distance between two multivariate Gaussian distributions and KID measures the dissimilarity between two probability distributions using samples drawn independently from each distribution. Lower FID and KID scores correspond to higher similarity between the real and synthetic images. We also evaluated synthetic images in terms of their distances from training images in the pixel space (Brock et al., 2019; Zhao et al., 2020). The main challenge in GAN training occurs when the generator D memorizes the input images rather than learning to generate new instances. We calculated the

L_1 distance between each synthetic image y and its three nearest neighbors in the training set, $NN_D(y)$. Images were first converted to gray-scale (pixel values in $[0, 255]$), and the Manhattan distance between image pairs was computed as the sum of absolute values in pixel intensities; the sum was normalized by the number of pixels and averaged among all generated images for each task and dataset.

Finally, synthesized images were evaluated in a blinded randomized test by two microsurgeons - one neurosurgeon and one otosurgeon- in terms of image authenticity and appearance. These experts first scored whether "this view was similar to what they would see under a microscope during training" in terms of tissues, instrument appearance (e.g. shadows cast by the instruments), and continuity of surgical sutures. Second, they assessed if they could "clearly detect and label the left- and right-hand tools in the image."

2.5. Downstream Task: Surgical Tool Detection

Finally, we examined whether the synthesized images improve tool tip detection. Total of 122 frames from the Practie-256-Train dataset was used to train the unconditional StyleGAN2+DiffAugment for 1044 kimg. Next, 1200 synthetic images were generated, of which 557 images with successful generation of left- and right hand surgical tools were selected. Tool tips were annotated with bounding boxes using LabelImg (Tzutalin, 2015). Finally, a YOLOv5-nano v.6.0 network (Jocher et al., 2021) was trained for 300 epochs five times separately using 1) the real-image dataset and 2) the combination of real- and synthetic-image dataset. Model hyperparameters are presented in Table S.2 in Appendix D. The input images were divided into training (n=102) and validation sets (n=20), and evaluated on a test set (n=488) using the unseen images from the Practice dataset. Mean average precision was compared at 0.5 threshold (mAP@0.5) from the two sets of five training runs with real and real+synthetic images.

3. Experiments and Results

Experiments were first conducted with 256-px samples to test the usefulness of incorporating class information (Section 3.1). The best architecture was selected for image synthesis at the 512-px level (Section 3.2). Appendix A provides the implementation details for each architecture and hyperparameter tuning. Table 2 summarizes all the results.

Table 2: FID and KID for Practice and INT datasets. Lower scores indicate more stable image generation.

	Experiment	**Knotting**		**Go-around**		**Neurosurg.**	
		FID	KID	FID	KID	FID	KID
256 px	Unconditional w tools	69.03	0.056	49.09	0.045	73.22	0.031
	Unconditional w and w/o tools	**61.25**	**0.042**	**42.16**	**0.034**	73.22	0.031
	Conditional w and w/o tools	67.94	0.051	68.01	0.065	-	-
512 px	Unconditional with and w/o tools	33.43	0.021	**25.17**	**0.021**	90.94	0.053

3.1. Low-Resolution Image Synthesis: Conditional and Unconditional Models

First, we experimented with DiffAugment and trained two unconditional models using 1) all the images and using 2) only the tool images. Initially, the conditional model resulted in low and highly unstable peformance during training as observed in the FID graph of the first 120 thousand images. The mini-batch size was adjusted so that the discriminator received samples with low variation and penalize or provide feedback to the generator to avoid the mode collapse (see Table S.1). Once training with DiffAugment achieved a stable GAN, each model was trained for 24 hours and the diversity and quality of synthesized images were evaluated using FID and KID. Figure S.1 in Appendix A demonstrates sample, without-tool images from conditional image generation experiments with the Knotting task.

Results in Table 2 indicate training the unconditional model with all the images resulted in generating more realistic images in comparison to the other two experiments for both practice tasks. While the conditional model performed slightly better than the unconditional with-tool model in Knotting, images generated for Go-around using the unconditional model were generally more diverse; images synthesized for Go-around demonstrated better diversity than the other two tasks using unconditional training.

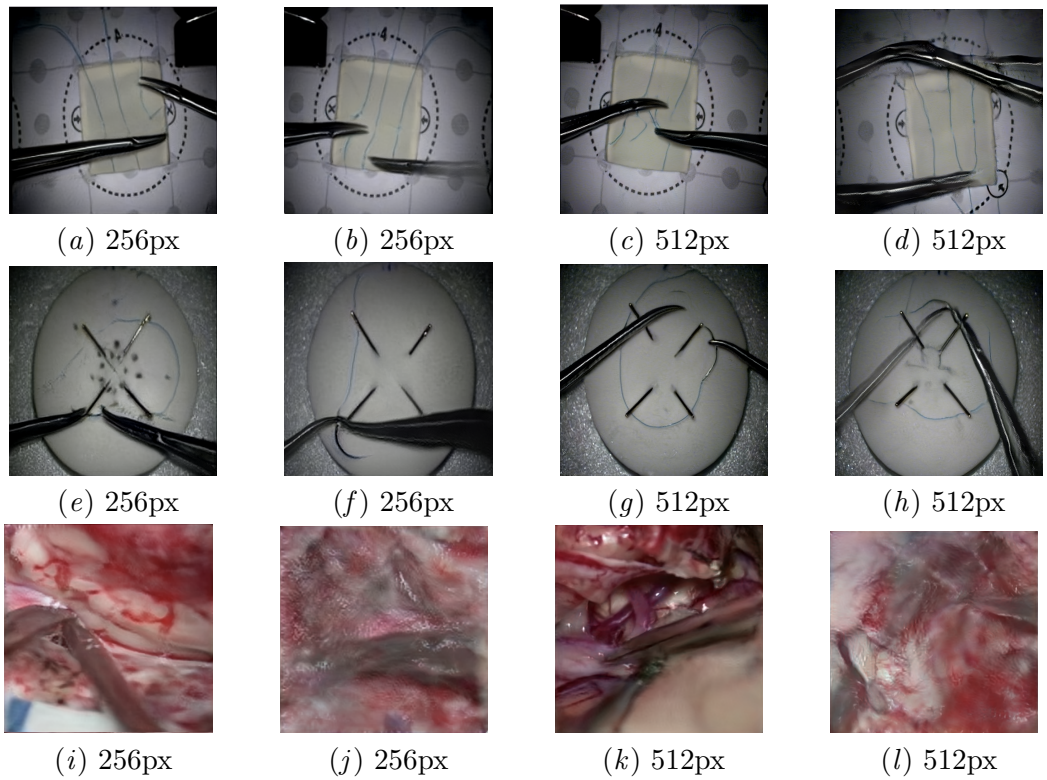
3.2. High-Resolution Image Synthesis: Unconditional Models

Considering the success of unconditional models with the low-resolution images, the differentiable augmentation model was separately trained using Practice-512-Train and INT-512-Train datasets. Final FID and KID values obtained from these experiments are presented in Table 2. These metrics generally improved for Practice dataset but show a deteriorated performance in the INT dataset. Images synthesized for Go-around received the best diversity scores. In perspective, FID scores of 42.16 and 25.17 are close to values reported for benchmarked DiffAug experiments with 1000 samples from LSUN-cat and FFHQ datasets (Zhao et al., 2020). Figure 2 presents sample synthetic images from low-resolution and high-resolution training sets and demonstrate successful examples as well as challenges such as partial tool generation and deformations.

Expert Evaluation A total of 602 512-px images generated by unconditional StyleGAN2+DiffAugment models from the Practice Dataset were submitted to two specialist microsurgeons for visual inspection. In response to question 1, on average, $40.03 \pm 34.53\%$ and $32.56 \pm 42.76\%$ of images generated for Knotting and Go-around, respectively, looked similar to what these experts would expect to observe under a surgical microscope. Figure S.2 in Appendix B includes eight sample images evaluated as realistic by both surgeons. Cohen's kappa, was below 0.02 for both tasks, showing these experts did not have similar expectations of synthetic images. To analyze responses to question 2 regarding the evaluation of individual surgical tools, Table 3 presents the number of synthesized images for high-resolution Practice datasets; the average detection of no-tool and correct labels are also presented. Tool detection was more successful for Knotting images and Cohen's kappa indicated a moderate level of agreement.

3.3. Nearest Neighbors Analysis in the Pixel Space

Manhattan distance was calculated between each synthetic image and nearest real images in the training sets of unconditional GAN experiments. Figure S.3 and Figure S.4 in Appendix C present images with the largest normalized L_1 distance from their 3 near-



(*a*) 256px (*b*) 256px (*c*) 512px (*d*) 512px

(*e*) 256px (*f*) 256px (*g*) 512px (*h*) 512px

(*i*) 256px (*j*) 256px (*k*) 512px (*l*) 512px

Figure 2: Synthetic images generated by unconditional GANs from low-resolution, and high-resolution images. The first and third column illustrate successful synthesis, while the images in the second and forth column show partial tool generation, deformation of tool tips and bodies, and unclear or blurry synthesis.

Table 3: Distribution of high-resolution images synthesized for Knotting and Go-around, and average ratio of correct detection from expert evaluations of surgical tools.

	Knotting		**Go-around**	
	Left-Hand	Right-Hand	Left-Hand	Right-Hand
With and Without Tool	(240, 61)	(249, 52)	(253, 48)	(250, 51)
Mean Detection	0.49 ± 0.03	0.44 ± 0.04	0.32 ± 0.03	0.32 ± 0.07
Kappa	0.70	0.61	0.58	0.65

est neighbors in different tasks and resolutions. Results of averaged normalized distances for each task and resolution in Table 4 indicate that GANs were, on average, more likely to generate new samples when trained by images of intraoperative procedures despite their challenging background and textures. The average, normalized distances were slightly larger when high-res images were used in the training sets. To provide a baseline for L_1 distances among real images, average values of 3NN-L_1 distances were computed. These distances were equal to 7.25 ± 3.63 (SE=0.13), 6.22 ± 2.97 (SE=0.11), and 28.82 ± 7.29 (SE=0.47) for Knotting, Go-around, and Neurosurgery training datasets. These within-real distances

Table 4: Mean and standard deviation for normalized L_1 distances between synthetic images and their three nearest neighbors. Standard errors are in the parentheses.

Dataset	Knotting	Go-around	Neurosurgery
256 px	18.98 ± 7.80 (0.55)	17.81 ± 7.21 (0.51)	25.29 ± 5.48 (0.78)
512 px	21.09 ± 8.07 (0.47)	19.68 ± 7.26 (0.42)	31.9 ± 7.11 (0.41)

are smaller than distances between the generated and real images in the first two Practice tasks but comparable to the normalized distances reported for the Intraoperative dataset.

3.4. Downstream task: Classification Results

Results from five runs showed the real training dataset achieved an average mAP@0.5 of 0.63 ± 0.03 and 0.68 ± 0.01 for left- and right-hand tools, respectively. After augmenting the training set with synthetic data from StyleGAN2+DiffAugment, average mAP@0.5 improved to 0.69 ± 0.02 and 0.68 ± 0.03. The improvement was statistically significant for the left-hand tool ($t = 3.82$, $p < 0.001$), but not for the right-hand tool ($t = 0.32$, $p = 0.75$). Mean mAP@0.5 for both hands improved by 0.03 ($t = 2.61$, $p = 0.03$).

4. Discussion and Conclusion

We present a novel work to investigate whether GANs can generate realistic images from limited datasets of microsurgical practice and intraoperative videos. We employed two variants of StyleGAN2, specifically designed for limited-data augmentation without mode collapse and training leakage. Experiments show that StyleGAN2+DiffAugment was more reliable than StyleGAN2+ADA in avoiding model collapse and generating realistic tool images. To investigate whether prior information about the tools in the image improves the generation results, we compared unconditional image synthesis with binary conditional GANs. Synthesis with 256-px datasets showed that using all training samples with unconditional models achieved better FID and KID scores.

Visual evaluations of 512-px images by two microsurgeons indicated a fidelity of over 30% for synthesized practice images. Different specializations and approaches in assessing regional features -–such as sharpness of background objects and tools and continuity of thin threads and background dots– could be key factors for this low agreement. However, they had moderate to high agreement in detecting and labeling left- and right-hand tools. Our analysis showed variations in the pose, shadows, and tool tip deformation in Go-around had hindered the detection of exact microinstruments. In a recent work, experts' performance was close to random in detecting real and fake histopathological images (Quiros et al., 2020). Existence of surgical instruments with well defined and familiar features in our images makes distinguishing the synthetic images easier. More investigation is needed to understand the nature and importance of visual features in expert assessment of surgical scenes.

Finally, our tool tip detection experiments using robust object detectors showed the feasibility of augmenting real images with synthesized data from unconditional GANs to improve tool tip detection in practice microsurgery datasets. Future work will include using transfer learning, exploring more variability in intraoperative images, and analyzing the latent space to verify diversity of synthetic image generation.

Acknowledgments

This study was partially funded by the Emil Aaltonen Foundation grant No. 64898 and Academy of Finland grants No. 334658 and No. 338492.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. doi: 10.5555/3305381.3305404.

Evgenii Belykh, Naomi R. Onaka, Irakliy T. Abramov, Kaan Yağmurlu, Vadim A. Byvaltsev, Robert F. Spetzler, Peter Nakaj, and Mark C. Preul. Systematic Review of Factors Influencing Surgical Performance: Practical Recommendations for Microsurgical Procedures in Neurosurgery. *World Neurosurg.*, 112:e182–e207, 2018. ISSN 18788769. doi: 10.1016/j.wneu.2018.01.005.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GaN training for high fidelity natural image synthesis. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–35, 2019. URL https://arxiv.org/abs/1809.11096.

Francesco Calimeri, Aldo Marzullo, Claudio Stamile, and Giorgio Terracina. Biomedical data augmentation using generative adversarial neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10614 LNCS(690974):626–634, 2017. ISSN 16113349. doi: 10.1007/978-3-319-68612-7_71.

François Chadebecq, Francisco Vasconcelos, Evangelos Mazomenos, and Danail Stoyanov. Computer Vision in the Surgical Operating Room. *Visceral Medicine*, 36(6):456–462, 2020. ISSN 2297475X. doi: 10.1159/000511934.

Joseph Davids, Savvas George Makariou, Hutan Ashrafian, Ara Darzi, Hani J. Marcus, and Stamatia Giannarou. Automated Vision-Based Microsurgical Skill Analysis in Neurosurgery Using Deep Learning: Development and Preclinical Validation. *World Neurosurgery*, 149:e669–e686, 2021. ISSN 18788769. doi: 10.1016/j.wneu.2021.01.117.

Shahram Eivazi, Roman Bednarik, Markku Tukiainen, Mikael Von Und Zu Fraunberg, Ville Leinonen, and Juha E. Jääskeläinen. Gaze behaviour of expert and novice microneurosurgeons differs during observations of tumor removal recordings. *Eye Track. Res. Appl. Symp.*, 1(212):377–380, 2012. doi: 10.1145/2168556.2168641.

Shahram Eivazi, Ahmad Hafez, Wolfgang Fuhl, Hoorieh Afkari, Enkelejda Kasneci, Martin Lehecka, and Roman Bednarik. Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta neurochirurgica*, 159(6): 959–966, 2017. ISSN 09420940. doi: 10.1007/s00701-017-3185-1.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. ISSN 15577317. doi: 10.1145/3422622.

Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, AlexWang1900, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu Diaconu, Mai Thanh Minh, Marc, albinxavi, fatih, oleg, and wanghaoyang0106. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support, October 2021. URL https://github.com/ultralytics/yolov5/tree/v6.0.

Takuhiro Kaneko, Kaoru Hiramatsu, and Kunio Kashino. Generative attribute controller with conditional filtered generative adversarial networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 7006–7015, 2017. doi: 10.1109/CVPR.2017.741.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00453.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Adv. Neural Inf. Process. Syst.*, 2020-Decem(NeurIPS), 2020a. ISSN 10495258.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020b. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00813.

Lauren R. Kennedy-Metz, Pietro Mascagni, Antonio Torralba, Roger D. Dias, Pietro Perona, Julie A. Shah, Nicolas Padoy, and Marco A. Zenati. Computer Vision in the Operating Room: Opportunities and Caveats. *IEEE Trans. Med. Robot. Bionics*, 3(1):2–10, 2021. ISSN 25763202. doi: 10.1109/TMRB.2020.3040002.

Jani Koskinen, Mastaneh Torkamani-Azar, Ahmed Hussein, Antti Huotarinen, and Roman Bednarik. Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery. *Computers in Biology and Medicine*, 141:105121, 2022. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.105121.

Tomi Leppänen, Hana Vrzakova, Roman Bednarik, Anssi Kanervisto, Antti-Pekka Elomaa, Antti Huotarinen, Piotr Bartczak, Mikael Fraunberg, and Juha E Jääskeläinen. Augmenting microsurgical training: Microsurgical instrument detection using convolutional neural networks. In *2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS)*, pages 211–216. IEEE, 2018. doi: 10.1109/CBMS.2018.00044.

Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical image analysis*, 76:102306, 2022. doi: 10.1016/j.media.2021.102306.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. URL https://arxiv.org/abs/1801.04406.

Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. pages 1–7, 2014. URL http://arxiv.org/abs/1411.1784.

Hiroki Morisako, Takeo Goto, Christian A Bohoun, Hironori Arima, Tsutomu Ichinose, and Kenji Ohata. Usefulness of the anterior transpetrosal approach for pontine cavernous malformations. *Neurosurgical Focus: Video*, 1(1):V4, 2019. doi: 10.3171/2019.7.FocusVid.19125. URL https://www.youtube.com/watch?v=2Q2CUhBbo28.

Markus Philipp, Anna Alperovich, Marielena Gutt-Will, Andrea Mathis, Stefan Saur, Andreas Raabe, and Franziska Mathis-Ullrich. Localizing neurosurgical instruments across domains and in the wild. In *Proceedings of Machine Learning Research*, volume 143, pages 581–595, 2021. URL https://2021.midl.io/proceedings/philipp21.pdf.

Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. PathologyGAN: Learning deep representations of cancer tissue. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121, pages 669–695. PMLR, 2020. URL https://proceedings.mlr.press/v121/quiros20a.html.

Pan Shi, Zijian Zhao, Sanyuan Hu, and Faliang Chang. Real-Time Surgical Tool Detection in Minimally Invasive Surgery Based on Attention-Guided Convolutional Neural Network. *IEEE Access*, 8:228853–228862, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3046258.

Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In *Int. Work. Simul. Synth. Med. imaging*, pages 1–11. Springer International Publishing, 2018. ISBN 9783030005351. doi: 10.1007/978-3-030-00536-8_1.

Ka Chun Siu, Irene H. Suh, Mukul Mukherjee, Dmitry Oleynikov, and Nick Stergiou. The impact of environmental noise on robot-assisted laparoscopic surgical performance. *Surgery*, 147(1):107–113, 2010. ISSN 00396060. doi: 10.1016/j.surg.2009.08.010.

Maribel Torres-Velazquez, Wei Jie Chen, Xue Li, and Alan B. McMillan. Application and Construction of Deep Learning Networks in Medical Imaging. *IEEE Trans. Radiat. Plasma Med. Sci.*, 5(2):137–159, 2021. ISSN 24697311. doi: 10.1109/TRPMS.2020.3030611.

Tzutalin. LabelImg. Git code, 2015. URL https://github.com/tzutalin/labelImg.

S. Swaroop Vedula and Gregory D. Hager. Surgical data science: The new knowledge domain. *Innovative surgical sciences*, 2(3):109–121, 2017. ISSN 23647485. doi: 10.1515/iss-2017-0004.

Yuta Yamazaki, Shingo Kanaji, Takeru Matsuda, Taro Oshikiri, Tetsu Nakamura, Satoshi Suzuki, Yuta Hiasa, Yoshito Otake, Yoshinobu Sato, and Yoshihiro Kakeji. Automated Surgical Instrument Detection from Laparoscopic Gastrectomy Video Images Using an Open Source Convolutional Neural Network Platform. *Journal of the American College of Surgeons*, 230(5):725–732.e1, 2020. ISSN 18791190. doi: 10.1016/j.jamcollsurg.2020.01.037.

Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33, 2020. URL https://arxiv.org/abs/2006.10738.

Appendix A. Image Synthesis: Implementation Details

Experiments for StyleGAN2 are run in TensorFlow and PyTorch. Images were preprocessed to assure all entries had the same square shapes and color space. Next, a folder of TFRecord transformed from images was generated for the purpose of efficient storage in TensorFlow version. For the Pytorch version with the class-conditional model, a metadata that contained label information was prepared, along with the uncompressed images before fitting the model. Codes used for image processing, creating labels, and calculating pixel-wise distances are publicly available at https://github.com/aprilycliu/GAN_toolkit.

Two open-source modules were utilized for running experiments in TensorFlow and PyTorch. Details of unconditional and conditional image synthesis with StyleGAN2+Ada are as follows. The Adaptive discriminator augmentation(ADA)[1] provides options to adjust augmentation setting on discriminator, including fixed or adaptively changed augmentation probability. From the diverse set of implemented transformations, we utilize the default `'Blit'`+ `'Geom'`+ `'Color'` options where blit refers to pixel blitting with x-flips, 90-degree rotations, and integer translation. Furthermore, for the Intraoperative Dataset, the horizontal mirror augmentation was enabled as it will double the training set and, in real life, surgeons may use their left hand as their dominant hand.

Lastly, the following transforms were used for unconditional and conditional image synthesis using StyleGAN2 with differentiable augmentation[2]: Translation within $[-1/8, 1/8]$ of the image size, padded with zeros; Cutout as masking with a random square of half image size; and Color, including random brightness within $[-0.5, 0.5]$, contrast within $[0.5, 1.5]$, and saturation within $[0, 2]$. In Zhao et al. (2020), the combination of Color + Translation + Cutout was especially effective and resulted in largest improvements from the baseline on CIFAR-10 benchmarks; this transformation was used in our experiments as well.

All experiments were conducted with two GPUs to ensure training stability. Training time was limited to 24 hours for each run. To control for model overfitting, the R_1 regularization term is applied to penalize the gradient in the discriminator D in case real data are generated by the generator G, i.e. $R_1(\psi) = \frac{\gamma}{2} E_{PD}(x) \left[\|\nabla D_\psi(x)\|^2\right]$ (Mescheder et al., 2018). The term ψ represents the discriminator weights, $E_{PD}(x)$ represents sampling from real samples, and γ represents a tunable hyperparameter. Other parameters included the mini-batch size of 32, learning rate of 0.001, R_1 regularization γ of 10, and a mapping net depth of 2. Table S.1 present the values of hyperparameters used for training the StyleGAN2-DiffAugment models. Networks were trained between 2200 and 2770 thousand images for generation of 256-px images. For generation of 512 images, models in tasks 1 and 2 were trained for 2760 kimg and the model of Neurosurgery for 3400 kimg.

Appendix B. Synthesized Images from Low- and High-Resolution Datasets

As discussed in Section 3.1, Figure S.1 demonstrates sample Knotting images generated using the conditional models of StyleGAN2+DiffAugment in the no-tool class.

Furthermore, as explained in Section 2.4, two microsurgeons were asked to evaluated 301 high-resolution images from Knotting and 301 high-resolution images from Go-around

1. https://github.com/NVlabs/stylegan2-ada
2. https://github.com/mit-han-lab/data-efficient-gans

Table S.1: Hyperparameters tuned for training unconditional and conditional GAN architectures. SD: Standard deviation; kimg: thousand trained images.

Parameter	**Knotting Uncond.**	**Knotting Cond.**	**Go-around Uncond.**	**Go-around Cond.**	**Neurosurgery Uncond.**
Learning rate	0.001	0.002	0.001	0.001	0.001
Mini-batch SD	4	4	4	8	4

tasks according to two criteria. Figure S.2 presents eight synthesized images that both experts evaluated as looking realistic or close to what they would expect to observe under a surgical microscope during a training session.

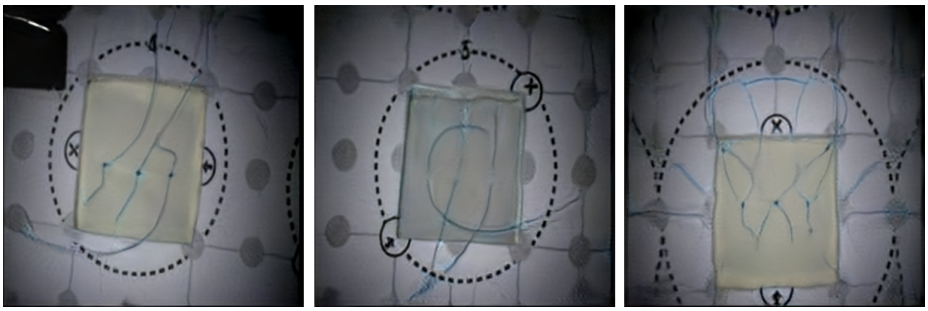


Figure S.1: Sample 256-px images generated by conditional models of StyleGAN2+DiffAugment from the no-tool class of the Knotting dataset. Numerical results of conducted experiments are presented in Table 2.

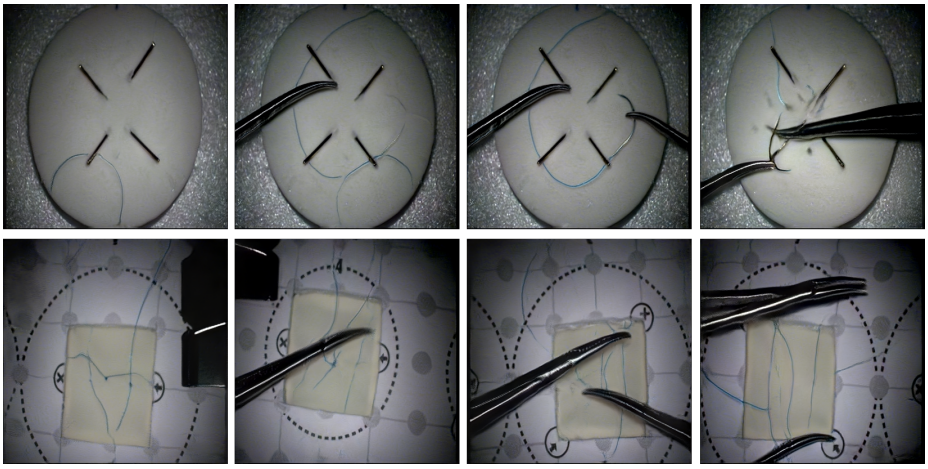


Figure S.2: Sample 512-px images from the Practice dataset that were unanimously evaluated as realistic by both experts.

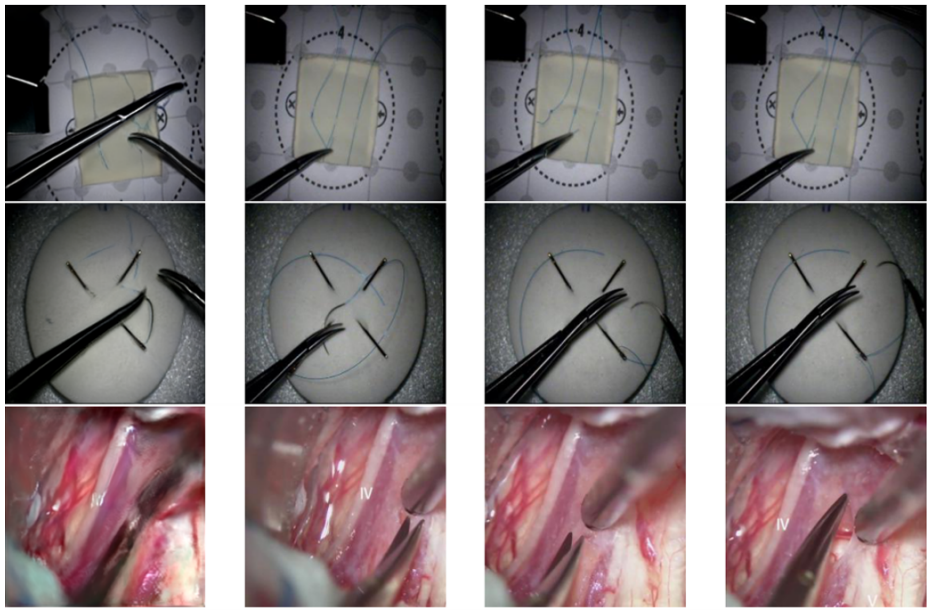


Figure S.3: Low-resolution images with largest normalized L_1 distances from their 3 nearest neighbors. Column one demonstrates sample synthetic images from the Knotting, Go-around, and Neurosurgery tasks. The three corresponding nearest neighbors are presented in columns two to four.

Appendix C. Presentation of Nearest Neighbors in Pixel Space

Figure S.3 and Figure S.4 demonstrate images with the largest normalized L_1 distance from their 3 nearest neighbors in different tasks and resolutions. In the case of Knotting and Go-around tasks in Figure S.3, the length and shape of tool tips has been modified but the poses still indicate a realistic use. In images synthesized from neurosurgery procedures in these figures, camera angles and tool poses are different from the original images used in the training sets.

Appendix D. Surgical Tool Detection: Implementation Details

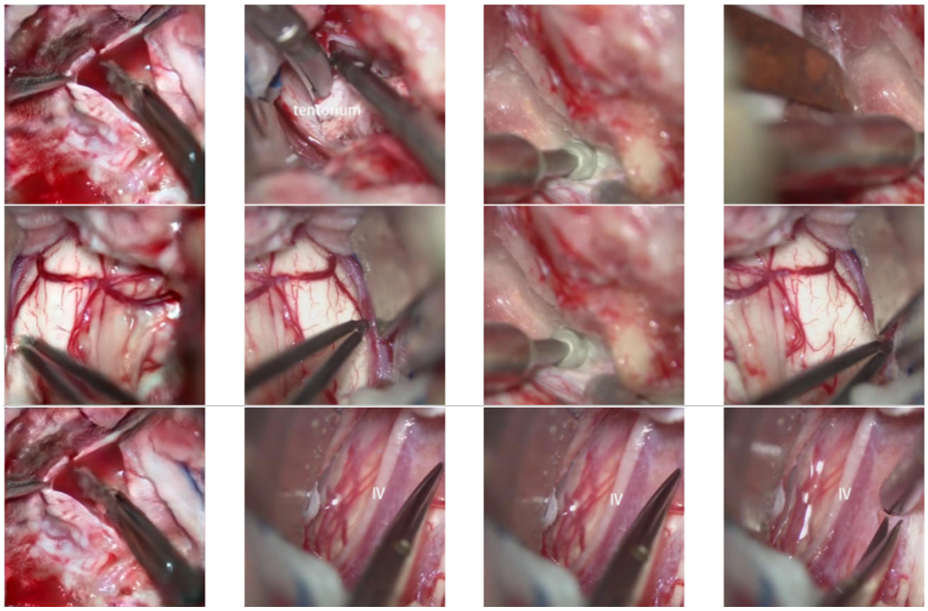


Figure S.4: High-resolution images with largest normalized L_1 distances from their 3 nearest neighbors. Column one demonstrates sample synthetic images from the Intraoperative Dataset and three nearest neighbors are presented in columns two to four.

Table S.2: Hyperparameters tuned for surgical tool detection using the YOLOv5-nano v6.0. network

Hyperparameter	Value	Hyperparameter	Value
lr0	0.01	fl_gamma	0
lrf	0.1	hsv_h	0.015
momentum	0.937	hsv_s	0.7
weight_decay	0.0005	hsv_v	0.4
warmup_epochs	3	degrees	0
warmup_momentum	0.8	translate	0.1
warmup_bias_lr	0.1	scale	0.5
box	0.05	shear	0
cls	0.5	perspective	0
cls_pw	1	flipud	0
obj	1	fliplr	0
obj_pw	1	mosaic	0
iou_t	0.2	mixup	0
anchor_t	4	copy_paste	0