Article

Development of a Deep Learning-Based Algorithm to Detect the Distal End of a Surgical Instrument

Hiroyuki Sugimori ¹ , Taku Sugiyama ^{2,*} , Naoki Nakayama ² , Akemi Yamashita ³ and Katsuhiko Ogasawara ¹

¹ Faculty of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan; sugimori@hs.hokudai.ac.jp (H.S.); oga@hs.hokudai.ac.jp (K.O.)

² Department of Neurosurgery, Hokkaido University Graduate School of Medicine, Sapporo 060-8638, Japan; naoki-na@med.hokudai.ac.jp

³ Graduate School of Health Sciences, Hokkaido University, Sapporo 060-0812, Japan; u950527@gmail.com

* Correspondence: takus1113@med.hokudai.ac.jp; Tel.: +81-11-706-5987

Received: 12 April 2020; Accepted: 18 June 2020; Published: 20 June 2020

Abstract: This work aims to develop an algorithm to detect the distal end of a surgical instrument using object detection with deep learning. We employed nine video recordings of carotid endarterectomies for training and testing. We obtained regions of interest (ROI; 32×32 pixels), at the end of the surgical instrument on the video images, as supervised data. We applied data augmentation to these ROIs. We employed a You Only Look Once Version 2 (YOLOv2)-based convolutional neural network as the network model for training. The detectors were validated to evaluate average detection precision. The proposed algorithm used the central coordinates of the bounding boxes predicted by YOLOv2. Using the test data, we calculated the detection rate. The average precision (AP) for the ROIs, without data augmentation, was 0.4272 ± 0.108 . The AP with data augmentation, of 0.7718 ± 0.0824 , was significantly higher than that without data augmentation. The detection rates, including the calculated coordinates of the center points in the centers of 8×8 pixels and 16×16 pixels, were 0.6100 ± 0.1014 and 0.9653 ± 0.0177 , respectively. We expect that the proposed algorithm will be efficient for the analysis of surgical records.

Keywords: object detection; deep learning; surgical video analysis

1. Introduction

In recent years, deep learning techniques have been adopted in several medical fields. The image classification [1–8], object detection [9–11] and semantic segmentation [12–17] techniques have been applied to diagnostic medical images, such as X-rays, computed tomography, ultrasound and magnetic resonance imaging, to support diagnosis and/or image acquisition. Since minimally invasive surgical procedures, such as robotic and endoscopic surgery, in which the operative field is viewed through an endoscopic camera, is an emerging field in medicine, surgical video-based analyses have also become a growing research topic, as they realize markerless and sensorless analysis without infringement of surgical safety regulations. These research approaches were used for the objective assessment of surgical proficiency, education, workflow optimization and risk management [18–20]. To date, several techniques have been proposed for the recognition and analysis of surgical instrument motion and behavior. Jo et al. [21] proposed a new real-time detection algorithm for surgical instrument detection using convolutional neural networks (CNNs). Zhao et al. [22] also proposed an algorithm for surgical instrument tracking, based on deep learning with line detection and a spatio-temporal context. A review of the literature [23] was recently reported for surgical tool detection studies. However, there have still been a few related works that deal with micro-neurosurgical or microvascular

specialties, because these surgeries require special surgical finesse and instruments, in which robotic and endoscopic means could not have been employed [24].

We previously reported “tissue motion” measurement results, as representative of the “gentle” handling of tissue during exposure of the carotid artery in a carotid endarterectomy (CEA). Therein, we employed off-the-shelf software for the video-based analysis [25]. However, we performed the analysis manually or semi-automatically; therefore, the analysis process was only retrospective, and quite time-consuming. Consequently, we could not realize real-time analysis. Based on these issues, we have introduced a deep learning technique for the analysis, because it is expected to realize automated analysis for real-time monitoring and feedback. Besides, for the analysis of surgical performance, a deep learning technique is expected to be more objective [18,19].

In this work, we focus on surgical instrument motion tracking during a CEA, as motion tracking information could be utilized to avoid surgical risk (e.g., unexpected tissue injury), as well as the objective assessment of surgical performance. For this purpose, we target the distal end of a surgical instrument, as this tip is the most important site that directly interacts with patient tissue (i.e., dissection, coagulation and grasping of the tissues) and surgical materials (i.e., holding of needles and cotton). Although there have been some studies that show the procedures for the detection of the shape or certain parts of surgical instruments (i.e., pose estimation of instruments) [21–24,26], we focus on the detection of the distal end as a goal in this work. This work aims to develop an algorithm to detect the distal end of a surgical instrument, using object detection with deep learning. We anticipate that the proposed algorithm can be used as an object detection technique with deep learning.

2. Materials and Methods

2.1. Subjects and Format of Video Records

We analyzed video records for nine patients who underwent a CEA, retrospectively. The video records were captured throughout the entire operation. The surgical video was recorded using a microscope-mounted video camera device. The video data was saved in the audio–video interlaced (AVI) format at 30 fps. Note that the institutional review board of Hokkaido University Hospital approved this work.

2.2. Preprocessing of Images

We preprocessed the dataset, as illustrated in Figure 1. The scene of the dataset was limited to the state in which the common, internal and external carotid arteries were revealed for the operation (Figure 2). The scene was determined by the logic that it would be the most likely to involve surgical risk during a CEA; the surgical video records contain many scenes, depending on the surgical procedures. We converted the AVI files to joint photographic experts group (JPEG) files. In total, we converted 512 images per patient to JPEG, from the full-length videos at 30 fps, to be able to detect the subtle motion of the surgical instrument. We cropped the video images of the shorter side to a square based on the longer side, because the AVI files were obtained using cameras from different vendors with different image ratios, i.e., 4:3 or 16:9. We padded the shorter side of the images with zeros, i.e., black. We performed these processes in-house with MATLAB software (The MathWorks, Inc., Natick, MA, USA).

2.3. Dataset

We created the supervised data to detect the end of the surgical instruments using in-house software (Figure 3). The software allowed the definition of an arbitrary 32×32 -pixel ROI at the end of the surgical instrument. The ROI data were output as a text file that included the object name and ROI coordinates. The object name was set to “tip” for one-class object detection. Typically, object names are assigned to several objects in multi-class detection. However, in this work, we have only one, since the detection target is only the end of the surgical instrument.

Figure 1. Procedure to select the scene and convert video to JPEG files for the dataset.

Figure 2. Representative images of the dataset for nine patients.

We divided the supervised data into nine subsets for nested cross-validation [27]. Even though there were several different ways to separate the dataset, we assigned the supervised data to training and test data. Each patient contributed 512 images; we used 4096 images from eight patients for training, and the remaining 512 images for testing (Figure 4). Each subset was an independent combination of eight patients for training and one patient for test images, to prevent the mixing of patient images between training and testing images within the subsets. We used the test dataset for the created model's evaluation, and not for the training process evaluation. To effectively learn for the

training dataset, we performed data augmentation [28–31] using image rotation from -90° to 90° in 5° steps, as illustrated in Figure 5.

Figure 3. The software for supervised data to detect the end of surgical instruments. The green circles indicate the end of a surgical instrument, while the yellow bounding boxes are the 32×32 -pixel ROIs.

Figure 4. Nine video records divided into nine subsets to complete nested cross-validation.

Figure 5. Data augmentation of the training datasets. The original images and supervised bounding boxes were rotated from -90° to 90° in 5° steps.

2.4. Training Images for Model Creation

We developed the software for object detection with a deep learning technique via in-house MATLAB software; we used a deep learning-optimized machine with an Nvidia Quadro P5000 graphics card (Nvidia Corporation, Santa Clara, CA, USA), 8.9 Tera floating-point single-precision operations per second, 288 GB/s memory bandwidth, and 16 GB memory per board. We performed the image training as transfer learning by initial weight using You Only Look Once Version 2 (YOLOv2) [32], with the MATLAB deep learning Toolbox and Computer Vision System Toolbox. The training model hyperparameters were as follows: maximum training epochs—10; initial learning rate—0.00001; mini-batch size—96. We used stochastic gradient descent with momentum for optimization with an initial learning rate. We set the momentum and L2 regulation to 0.9 and 0.0001, respectively. We performed image training nine times based on the training subsets in Figure 2.

2.5. Evaluation of Created Models

We incorporated the predicted bounding boxes into the MATLAB software, to reveal the region of the distal end of the surgical instrument as a bounding box. We also evaluated the detection of the region of the distal end using average precision (AP), log-average miss rate (LAMR), and frame per second (FPS) for the efficiency of the created model. We examined the bounding boxes based on the supervised ROI, according to the “evaluateDetectionPrecision” and “evaluateDetectionMissRate” function in the MATLAB Computer Vision Toolbox.

2.6. Algorithm and Evaluation of Distal End Detection

The proposed algorithm to detect the distal end of a surgical instrument used the central coordinates of the predicted bounding boxes, and was completed with YOLOv2. The predicted bounding box had a square shape around the center point of the distal end of the surgical instrument. Therefore, we can use coordinate information from the center of the bounding box as a mark to display the distal end of the surgical instrument. The evaluation of whether the calculated coordinates of the center point were normally displayed was confirmed in order to compare the position of the supervised coordinates. The detection rate of the distal end of the surgical instrument was defined as including the calculated coordinates of the center point of 8×8 or 16×16 pixels among the valid bounding boxes (Figure 6). We calculated the detection rate of the distal end using the test data of 512 image. We calculated these processes with MATLAB software. We presented all the results as mean and standard deviation (SD) according to the number of nine-fold subsets.

Figure 6. Process to calculate the detection rate of the distal end of the surgical instrument.

2.7. Statistical Analysis

We presented AP, LAMR, FPS, and the detection rate of the distal end of the surgical instrument as mean $\pm$ SD. We evaluated the comparison of the effect with and without data augmentation using the Mann–Whitney U-test at the 0.05 significance level.

3. Results

3.1. Bounding Box Detection

Table 1 summarizes the AP for the 32×32 -pixel ROIs detected as a bounding box trained by YOLOv2. The AP, LAMR and FPS for the 32×32 -pixel ROIs with data augmentation are 0.4272 ± 0.108 , 0.6619 ± 0.0703 and 43.3 ± 0.8 , respectively, while with data augmentation, they are 0.7718 ± 0.0824 , 0.3488 ± 0.1036 and 29.9 ± 4.5 , respectively. These results with data augmentation are significantly higher than without data augmentation. Figure 7 illustrates representative examples of the bounding box detection results.

Table 1. Average precision (AP), log-average miss rate (LAMR), and frame per second (FPS) for 32×32 -pixel ROIs.

	Without Data Augmentation			With Data Augmentation		
	AP	LAMR	FPS [fps]	AP	LAMR	FPS [fps]
Subset A	0.3895	0.7275	43.2	0.8258	0.4234	41.9
Subset B	0.4649	0.6361	43.5	0.6929	0.4382	28.0
Subset C	0.4958	0.6216	44.0	0.6866	0.4206	28.3
Subset D	0.2960	0.7263	44.1	0.6689	0.4124	28.0
Subset E	0.4150	0.6515	43.9	0.8639	0.2006	28.1
Subset F	0.3104	0.7367	43.8	0.8533	0.4117	28.7
Subset G	0.6216	0.5288	43.4	0.8603	0.1821	28.3
Subset H	0.3336	0.7159	41.5	0.8055	0.3841	28.7
Subset I	0.5179	0.6124	42.7	0.9083	0.2585	28.7
mean $\pm$ SD	0.4272 ± 0.108	0.6619 ± 0.0703	43.3 ± 0.8	0.7962 ± 0.0897	0.3488 ± 0.1036	29.9 ± 4.5

3.2. Detection Rate of the Distal End of the Surgical Instrument

Table 2 summarizes the detection rate of the distal end of the surgical instrument. The detection rates, including the calculated coordinates of the center point of the 8×8 and 16×16 pixels, are 0.6100 ± 0.1014 and 0.9653 ± 0.0177 , respectively. Figure 7 illustrates representative examples of results for the detection of the distal end of the surgical instrument.

Figure 7. Representative examples of the bounding box and surgical instrument distal end detection results.

Table 2. The detection rate of the distal end of the surgical instrument.

	Detection Rate of the Distal End of the Surgical Instrument	
	Within the Center of 8×8 Pixels	Within the Center of 16×16 Pixels
Subset A	0.5388	0.9660
Subset B	0.4916	0.9241
Subset C	0.5826	0.9604
Subset D	0.5033	0.9560
Subset E	0.6430	0.9706
Subset F	0.6742	0.9796
Subset G	0.8210	0.9828
Subset H	0.6419	0.9733
Subset I	0.5934	0.9750
mean $\pm$ SD	0.6100 ± 0.1014	0.9653 ± 0.0177

4. Discussion

The purpose of this work was to develop and evaluate an algorithm to detect the distal end of a surgical instrument, using object detection with deep learning, during surgery. With regard to detection as a bounding box, the AP with data augmentation was 0.7718 ± 0.0824 . This result confirmed that the data augmentation was able to achieve less dependency on location and angle of the distal end of surgical instruments, even though such instruments have a variety of types and shapes. We set the rotation angle of the data augmentation to -90° – 90° , in 5° steps, because the view of the surgical field was commonly recorded with reference to the view of the main surgeon. Approximately 80% of the AP was appropriate for comparing our results with those of another report [9] on the detection of small anatomies of the brain with deep learning. The FPS with data augmentation showed a slight variation, because the surgical instrument's motion was relatively different throughout the images of the test data, depending on the surgical procedure of that scene. The relationships between the position or motion of the detectable object, and the precision or FPS of the detection, have been reported [33,34], even though the entire surgical instrument's motions through the test data were not evaluated. With regard to the proposed algorithm for the detection of the distal end of the surgical instrument, the detection rates for the center point of the 8×8 and 16×16 pixels were 0.6100 ± 0.1014 and 0.9653 ± 0.0177 , respectively, even though the evaluated condition was limited to valid bounding boxes. Similarly to previous results for the assessment of surgical skill level in the surgical operation [18,19], our results

indicate that the proposed algorithm is efficient as an indicator for the analysis of a surgical operation. The proposed algorithm for the detection of the distal end of the surgical instrument as a point differed procedurally from those in previous reports [35,36], in which the proposed method was calculated using the center coordinates of the bounding box. CNN architectures for object detection have been widely used with many deep learning frameworks [32,37–39]. The advantage of this algorithm is that it creates the dataset using the existing YOLOv2 [32]. Specifically, the only attempt to define the square size of the ROIs around the distal end of the surgical instruments that was easy was that which did not require the information of the whole instrument, because the surgical instrument states varied in the operation fields of the video record. For the distal end detection, the calculation of the coordinates in the image that depicted the distal end of the surgical instruments was not necessary to complete the calculation, because the proposed bounding boxes always showed as a square shape.

The limitations of the present work are as follows. First, the scenes of the surgical record were limited to CEAs. Although the surgical procedures were performed in many parts of the body, in a preliminary study, it was necessary to focus on a single procedure. There are many types of surgical instruments, such as monopolar and bipolar handpieces, biological tweezers, forceps, surgical scissors and scalpels. Furthermore, the evaluation of a surgical instrument detection method has been reported [23], with many procedures and datasets. Therefore, depending on the various types of operations and techniques, a direct comparison of surgical instruments, under the same conditions, should be considered in future work. Second, the present work only focused on the scene where carotid arteries were exposed. The primary reason for focusing on this scene was to evaluate a delicate operative technique with inherent risks, such as cranial nerve injury, the possibility of ischemic stroke caused by plaque disruption, and hemorrhage. This scene was known to involve the most technical surgical performances; that is, those during a CEA [40]. Third, with regard to the conversion from surgical records to JPEG files, the training data in this work used a 30 fps temporal resolution for all conversions. In this work, the scene of the dataset was selected in order to focus on detecting the feature of the instruments via the time resolution, rather than letting the amount of features obtained depend on the length of time. Reducing the frame rate for JPEG conversion would provide more image features; however, the repeating pattern of similar movements might not be affected by the improvement associated with obtaining new image features. The fps of the conversion was adequate for the given procedure, because the approach to the blood vessels was a particularly important part of the surgery, involving the potential risk of bleeding. Nevertheless, this approach contributed to these novel findings and insights in the field of the evaluation of the method of detecting the distal end of the surgical instrument. Note that previous works [41–43] have reported the detection of the point of the joint for animals and human body parts. However, the approach reported in these works required a dedicated technique to train the dataset. Moreover, many works [35,36,44,45] have reported the detection of entire surgical instruments using semantic segmentation. These works were used in the Endoscopic Vision Challenge (<http://endovis.grand-challenge.org>) dataset at the international conference on Medical Image Computing and Computer Assisted Intervention. To the best of our knowledge, detection of the distal end of a surgical instrument, with a particular focus on CEA operations, has not been reported, and this we have done, even though our work was based on a currently limited dataset. The specific target surgery uses similar techniques and instruments, even though the operations have a variety of purposes and employ a variety of techniques. As with all common procedures, the instruments were almost the same for all operations. In particular, the scene of the approach to the vein focuses on the vessels and the instruments. Therefore, the detection of the distal end of the surgical instrument with deep learning will be useful for common surgical procedures.

5. Conclusions

We performed this research to develop an algorithm to detect the distal end of a surgical instrument in CEA operations, using object detection with deep learning. We determined that the detection of the distal end of the surgical instrument attained a high AP and detection rate. Our proposed

algorithm could be achieved as one of the real-time detection algorithms for the detection of surgical instruments using CNNs. This algorithm has the potential to improve the objective assessment of surgical proficiency, education, workflow optimization and risk management. Based on this algorithm, we will conduct future research, with a larger case-cohort, that analyzes the correlation between the surgical instrument's motion and the surgeon skill level and/or clinical results (e.g., adverse events in surgery). The use of the proposed algorithm in this work could be helpful for surgical record analysis.

Author Contributions: H.S. contributed to data analysis, algorithm construction, and the writing and editing of the article. T.S. and N.N. proposed the idea and contributed to data acquisition, and reviewed and edited the paper. A.Y. also contributed to reviewing and editing. K.O. performed supervision and project administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors thank the Ogasawara's laboratory research students Hang Ziheng and Han Feng for their help.

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

References

1. Sugimori, H. Classification of computed tomography images in different slice positions using deep learning. *J. Healthc. Eng.* **2018**, *2018*, 1753480. [[CrossRef](#)] [[PubMed](#)]
2. Sugimori, H. Evaluating the overall accuracy of additional learning and automatic classification system for CT images. *Appl. Sci.* **2019**, *9*, 682. [[CrossRef](#)]
3. Swati, Z.N.K.; Zhao, Q.; Kabir, M.; Ali, F.; Ali, Z.; Ahmed, S.; Lu, J. Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput. Med. Imaging Graph.* **2019**, *75*, 34–46. [[CrossRef](#)] [[PubMed](#)]
4. Yasaka, K.; Akai, H.; Kunimatsu, A.; Abe, O.; Kiryu, S. Liver fibrosis: Deep convolutional neural network for staging by using gadoteric acid-enhanced hepatobiliary phase MR images. *Radiology* **2017**, *287*, 146–155. [[CrossRef](#)]
5. Zhang, Q.; Ruan, G.; Yang, W.; Liu, Y.; Zhao, K.; Feng, Q.; Chen, W.; Wu, E.X.; Feng, Y. MRI Gibbs-ringing artifact reduction by means of machine learning using convolutional neural networks. *Magn. Reson. Med.* **2019**, *82*, 2133–2145. [[CrossRef](#)]
6. Roth, H.R.; Lee, C.T.; Shin, H.-C.; Seff, A.; Kim, L.; Yao, J.; Lu, L.; Summers, R.M. Anatomy-specific classification of medical images using deep convolutional nets. In Proceedings of the 2015 IEEE International Symposium on Biomedical Imaging, New York, NY, USA, 16–19 April 2015; pp. 101–104.
7. Noguchi, T.; Higa, D.; Asada, T.; Kawata, Y.; Machitori, A.; Shida, Y.; Okafuji, T.; Yokoyama, K.; Uchiyama, F.; Tajima, T. Artificial intelligence using neural network architecture for radiology (AINNAR): Classification of MR imaging sequences. *Jpn. J. Radiol.* **2018**, *36*, 691–697. [[CrossRef](#)]
8. Gao, X.W.; Hui, R.; Tian, Z. Classification of CT brain images based on deep learning networks. *Comput. Methods Programs Biomed.* **2017**, *138*, 49–56. [[CrossRef](#)]
9. Sugimori, H.; Kawakami, M. Automatic detection of a standard line for brain magnetic resonance imaging using deep learning. *Appl. Sci.* **2019**, *9*, 3849. [[CrossRef](#)]
10. Cao, Z.; Duan, L.; Yang, G.; Yue, T.; Chen, Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC Med. Imaging* **2019**, *19*, 1–9. [[CrossRef](#)]
11. Arijji, Y.; Yanashita, Y.; Kutsuna, S.; Muramatsu, C.; Fukuda, M.; Kise, Y.; Nozawa, M.; Kuwada, C.; Fujita, H.; Katsumata, A.; et al. Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique. *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **2019**, *128*, 424–430. [[CrossRef](#)]
12. Ge, C.; Gu, I.Y.H.; Jakola, A.S.; Yang, J. Deep learning and multi-sensor fusion for glioma classification using multistream 2D convolutional networks. In Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, HI, USA, 17–22 July 2018; Volume 2018, pp. 5894–5897.

13. Dalmiş, M.U.; Litjens, G.; Holland, K.; Setio, A.; Mann, R.; Karssemeijer, N.; Gubern-Mérida, A. Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Med. Phys.* **2017**, *44*, 533–546. [[CrossRef](#)] [[PubMed](#)]
14. Duong, M.T.; Rudie, J.D.; Wang, J.; Xie, L.; Mohan, S.; Gee, J.C.; Rauschecker, A.M. Convolutional neural network for automated FLAIR lesion segmentation on clinical brain MR imaging. *Am. J. Neuroradiol.* **2019**, *40*, 1282–1290. [[CrossRef](#)] [[PubMed](#)]
15. Havaei, M.; Davy, A.; Warde-Farley, D.; Biard, A.; Courville, A.; Bengio, Y.; Pal, C.; Jodoin, P.M.; Larochelle, H. Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **2017**, *35*, 18–31. [[CrossRef](#)] [[PubMed](#)]
16. Chen, L.; Bentley, P.; Rueckert, D. Fully automatic acute ischemic lesion segmentation in DWI using convolutional neural networks. *NeuroImage Clin.* **2017**, *15*, 633–643. [[CrossRef](#)] [[PubMed](#)]
17. Khalili, N.; Lessmann, N.; Turk, E.; Claessens, N.; de Heus, R.; Kolk, T.; Viergever, M.A.; Benders, M.J.N.L.; Išgum, I. Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magn. Reson. Imaging* **2019**, *64*, 77–89. [[CrossRef](#)] [[PubMed](#)]
18. Sugiyama, T.; Lama, S.; Gan, L.S.; Maddahi, Y.; Zareinia, K.; Sutherland, G.R. Forces of tool-tissue interaction to assess surgical skill level. *JAMA Surg.* **2018**, *153*, 234–242. [[CrossRef](#)]
19. Birkmeyer, J.D.; Finks, J.F.; O'Reilly, A.; Oerline, M.; Carlin, A.M.; Nunn, A.R.; Dimick, J.; Banerjee, M.; Birkmeyer, N.J.O. Surgical skill and complication rates after bariatric surgery. *N. Engl. J. Med.* **2013**, *369*, 1434–1442. [[CrossRef](#)]
20. Elek, R.N.; Haidegger, T. Robot-assisted minimally invasive surgical skill assessment—Manual and automated platforms. *Acta Polytech. Hung.* **2019**, *16*, 141–169.
21. Jo, K.; Choi, Y.; Choi, J.; Chung, J.W. Robust real-time detection of laparoscopic instruments in robot surgery using convolutional neural networks with motion vector prediction. *Appl. Sci.* **2019**, *9*, 2865. [[CrossRef](#)]
22. Zhao, Z.; Cai, T.; Chang, F.; Cheng, X. Real-time surgical instrument detection in robot-assisted surgery using a convolutional neural network cascade. *Healthc. Technol. Lett.* **2019**, *6*, 275–279. [[CrossRef](#)]
23. Bouget, D.; Allan, M.; Stoyanov, D.; Jannin, P. Vision-based and marker-less surgical tool detection and tracking: A review of the literature. *Med. Image Anal.* **2017**, *35*, 633–654. [[CrossRef](#)] [[PubMed](#)]
24. Bouget, D.; Benenson, R.; Omran, M.; Riffaud, L.; Schiele, B.; Jannin, P. Detecting surgical tools by modelling local appearance and global shape. *IEEE Trans. Med. Imaging* **2015**, *34*, 2603–2617. [[CrossRef](#)] [[PubMed](#)]
25. Sugiyama, T.; Nakamura, T.; Ito, Y.; Tokairin, K.; Kazumata, K.; Nakayama, N.; Houkin, K. A pilot study on measuring tissue motion during carotid surgery using video-based analyses for the objective assessment of surgical performance. *World J. Surg.* **2019**, *43*, 2309–2319. [[CrossRef](#)] [[PubMed](#)]
26. Zhao, Z.; Chen, Z.; Voros, S.; Cheng, X. Real-time tracking of surgical instruments based on spatio-temporal context and deep learning. *Comput. Assist. Surg.* **2019**, *24*, 20–29. [[CrossRef](#)] [[PubMed](#)]
27. Vabalas, A.; Gowen, E.; Poliakoff, E.; Casson, A.J. Machine learning algorithm validation with a limited sample size. *PLoS ONE* **2019**, *14*, 1–20. [[CrossRef](#)]
28. Mikołajczyk, A.; Grochowski, M. Data augmentation for improving deep learning in image classification problem. In Proceedings of the 2018 International Interdisciplinary Ph.D. Workshop (IIPhDW), Świnoujście, Poland, 9–12 May 2018; pp. 117–122.
29. Mash, R.; Borghetti, B.; Pecarina, J. Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks. In Proceedings of the Advances in Visual Computing; Springer International Publishing: Cham, Switzerland, 2016; pp. 113–122.
30. Taylor, L.; Nitschke, G. Improving deep learning using generic data augmentation. *arXiv* **2017**, arXiv:1708.06020. Available online: <https://arxiv.org/abs/1708.06020> (accessed on 1 April 2020).
31. Eaton-rosen, Z.; Bragman, F. Improving data augmentation for medical image segmentation. In Proceedings of the 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands, 4–6 July 2018; pp. 1–3.
32. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
33. Lalonde, R.; Zhang, D.; Shah, M. ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4003–4012.

34. Chen, K.; Wang, J.; Yang, S.; Zhang, X.; Xiong, Y.; Loy, C.C.; Lin, D. Optimizing video object detection via a scale-time lattice. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7814–7823.
35. Hasan, S.M.K.; Linte, C.A. U-NetPlus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. In Proceedings of the 41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 7205–7211.
36. Ni, Z.-L.; Bian, G.-B.; Xie, X.-L.; Hou, Z.-G.; Zhou, X.-H.; Zhou, Y.-J. RASNet: Segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In Proceedings of the 41th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Berlin, Germany, 23–27 July 2019; pp. 5735–5738.
37. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2014, Columbus, OH, USA, 24–27 June 2014; pp. 580–587.
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
39. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
40. Oikawa, K.; Kato, T.; Oura, K.; Narumi, S.; Sasaki, M.; Fujiwara, S.; Kobayashi, M.; Matsumoto, Y.; Nomura, J.i.; Yoshida, K.; et al. Preoperative cervical carotid artery contrast-enhanced ultrasound findings are associated with development of microembolic signals on transcranial Doppler during carotid exposure in endarterectomy. *Atherosclerosis* **2017**, *260*, 87–93. [[CrossRef](#)]
41. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1302–1310.
42. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W.; Bethge, M. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **2018**, *21*, 1281–1289. [[CrossRef](#)]
43. Papandreou, G.; Zhu, T.; Chen, L.C.; Gidaris, S.; Tompson, J.; Murphy, K. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 282–299.
44. Allan, M.; Shvets, A.; Kurmann, T.; Zhang, Z.; Duggal, R.; Su, Y.-H.; Rieke, N.; Laina, I.; Kalavakonda, N.; Bodenstedt, S.; et al. 2017 robotic instrument segmentation challenge. *arXiv* **2019**, arXiv:1902.06426. Available online: <https://arxiv.org/abs/1902.06426> (accessed on 1 April 2020).
45. Bernal, J.; Tajikbakhsh, N.; Sanchez, F.J.; Matuszewski, B.J.; Chen, H.; Yu, L.; Angermann, Q.; Romain, O.; Rustad, B.; Balasingham, I.; et al. Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge. *IEEE Trans. Med. Imaging* **2017**, *36*, 1231–1249. [[CrossRef](#)] [[PubMed](#)]