applied sciences

MDPI

Article

Automatic Surgical Instrument Recognition—A Case of Comparison Study between the Faster R-CNN, Mask R-CNN, and Single-Shot Multi-Box Detectors

Jiann-Der Lee [1,2,3], Jong-Chih Chien [4,*], Yu-Tsung Hsu [1] and Chieh-Tsai Wu [2]

1 Department of Electrial Engineering, Chang Gung University, Taoyuan 33302, Taiwan; jdlee@mail.cgu.edu.tw (J.-D.L.); daifwcou@gmail.com (Y.-T.H.)
2 Department of Neurosurgery, Chang Gung Memorial Hospital at Linkou, Taoyuan 33305, Taiwan; woodie2@adm.cgmh.org.tw
3 Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 243303, Taiwan
4 School of Informatics, Kainan University, Taoyuan 33857, Taiwan
* Correspondence: jcchien@gapps.knu.edu.tw; Tel.: +886-3-341-2500 (ext. 6191)

Abstract: In various studies, problems with surgical instruments in the operating room are usually one of the major causes of delays and errors. It would be of great help, in surgery, to quickly and automatically identify and keep count of the surgical instruments in the operating room using only video information. In this study, the recognition rate of fourteen surgical instruments is studied using the Faster R-CNN, Mask R-CNN, and Single Shot Multi-Box Detectors, which are three deep learning networks in recent studies that exhibited near real-time object detection and identification performance. In our experimental studies using screen captures of real surgery video clips for training and testing, this study found that that acceptable accuracy and speed tradeoffs can be achieved by the Mask R-CNN classifier, which exhibited an overall average precision of 98.94% for all the instruments.

Keywords: deep learning networks; surgical instruments

Citation: Lee, J.-D.; Chien, J.-C.; Hsu, Y.-T.; Wu, C.-T. Automatic Surgical Instrument Recognition—A Case of Comparison Study between the Faster R-CNN, Mask R-CNN, and Single-Shot Multi-Box Detectors. *Appl. Sci.* **2021**, *11*, 8097. https://doi.org/10.3390/app11178097

Academic Editor: Flavio Cannavò

Received: 27 June 2021
Accepted: 29 August 2021
Published: 31 August 2021

1. Introduction

In various studies on equipment-related incidents in the operating room, it was found that equipment-related type of error occurs in about 15.9% of cases involved in the studies [1,2], mostly due to unavailability of the requested instruments, which caused significant delays in the surgery. Mistakes by assisting nurses in identifying the correct instrument will be minimized if the location of the correct instrument can be automatically identified. It would be helpful to assisting nurses as well as the surgeons if a detector can be used to identify and track the surgical equipment before, during, and after the surgery in a near real-time manner using just the videos available during the surgery. This type of detector could help reduce incidents such as accidentally leaving surgical instruments inside of patients. There have been proposals in the literature to embed RFID (Radio Frequency Identification) tags onto surgical instrument for tracking [3], build special apparatus to help track surgical instruments [4], or use machine-learning features without deep learning to distinguish between groups of instruments, such as in [5], in which the attention-HOG (Histogram of Gradients) feature was used to distinguish between five types of instruments, and it reported an average accuracy of 90.1%. There have also been studies using modified deep learning networks on medical image processing datasets provided by various medical image processing challenge events, such as Cai et al. [6], whose paper reported a modified CNN architecture with an accuracy of 75% (overall) on the EndoVis challenge dataset [7], which had 15 different surgical instruments, and a 92.1% accuracy on their own dataset. A paper presented by Zhao et al. [8] shows that their modified CNN architecture achieved a 91.6% mean average precision accuracy on the Atlas Dione dataset, containing a specified number of surgical instruments, which is just

slightly better than the performance of the unmodified Faster R-CNN [9] network reported by the same study. Yu et al. [10] proposed a modified SSD (Single-Shot Multi-Box Detector) classifier for real-time processing of videos containing surgical instruments, and it reported that the modified architecture achieved an average precision of 90.08% in a set of images extracted from real surgery video clips.

Classification is a combined discipline of supervised machine learning and statistics classification to group (or order) groups of objects with similarities. The features used in determining the similarities between groups can be extracted from the objects using machine learning methods, and in the CNN-based deep learning architectures, these features can be extracted by performing consecutive convolutions; more details may be extracted by increasing the number of layers of convolution. Then, the areas surrounding the objects of interest (called regions of interest, or ROI) are computed from these features. In most cases, the training and testing sets may be altered using different settings for re-testing. Then, the operations of statistics are used to compute the similarities between the features then classify similar objects together. In supervised learning, the objects to be classified are separated into a pre-labeled training set and testing set, which are treated as unlabeled during testing. The training set is used to train the deep learning network so that it would classify objects in this set correctly while reducing classification errors in successive iterations until a threshold is achieved. Then, the trained network is tested using the testing set to classify them accordingly; then, the accuracy, precision, and recall are computed based on the classified results against their hidden labels. If the results of the testing are unsatisfactory, then additional trainings or increasing the size of the training set may be implemented. In a multi-class classification problem, the mean average precision (MAP) for multiple testing should be sufficient to determine the "goodness" of the classifier for the objects to be classified. Some only report the average MAP for all the classes, but reporting the MAP for each class would be more helpful in understanding the performance of the classifier.

Previous studies using hardware-based solutions, such as RFID, tend to be more stable, but it requires additional financial investment as well as possibly increases the size of the instruments, making them less user-friendly. Machine learning features such as attention-HOG can be selected by the designer and offer more explainability of the grouping results, but they tend to require more computationally intensive pre-processing for each image or video frame, and they are less likely to be able to achieve near real-time performance during classification. The deep-learning architectures seem to be a possible solution to achieve real-time classification, since the feature to be extracted can be computed in real time. However, the detection and classification of the objects can be a bottleneck, especially if the objects to be classified in different classes may exhibit highly similar features. Several modified deep-learning architectures have been proposed in the literature, as discussed above. However, it is not the aim of this paper to propose a new architecture nor is it to compare with the results of previous publications, since the datasets they used are not publically available now. The aim of this paper is to investigate a few well-known, non-modified, deep learning architectures, and based on their MAP performances in classifying the surgical instruments used in a local hospital, choose it as the base model for modification in continuing the investigation in the future.

Previous studies presented their results in MAP (mean average precision) of all the surgical instruments as a whole without reporting on the accuracy of each instrument, which is not useful for understanding their performances on each individual type of instrument and the instrument or instruments most likely to be misidentified. So, the purpose of this paper is to investigate the recognition performances on surgical instruments without building a new system but by choosing a few readily available learning architectures that have been studied in other applications without modification to their architecture. The reason for this is that the designs of deep learning architectures are flexible and can be modified to achieve better performances, so it would be important to identify a good base architecture for the purpose of detection and identification of medical instruments, and

then extend and modify this base architecture in the future in order to achieve better performances. The purpose of this study is a preliminary investigation into which base deep learning architecture is most suited for classifying surgical instruments. In order to achieve this objective, this study selects three basic deep learning architectures, experimentally finds their accuracies on detecting and identifying individual surgical instruments in our dataset, and compares their performances. The dataset used in this study was generated by extracting frames from real surgery video clips provided by the neurosurgery department of a local hospital in Taiwan. Based on the results of the study presented in [8], the Faster R-CNN (Region-based Convolutional Neural Networks) classifier is chosen as one of the base architectures to study. In addition, the Mask R-CNN [11] classifier is also chosen, since it was claimed that the ROI (region-of-interest)-pooling operation used by the Faster R-CNN network is less accurate and may adversely affect its classification accuracy. So, the Mask R-CNN network proposed a replacement stage, the ROI-alignment stage, in order to get a better accuracy in bounding boxes around the object to be classified. Since the base architecture for the Mask R-CNN network can be considered as slightly different from the Faster R-CNN network, it is also included in this study. In addition, various other studies in object detection [12,13] showed in their results that the SSD (Single-Shot Multi-Box Detector) [14] network is another near real-time architecture that shows an average accuracy performance similar to the Faster R-CNN network for detecting common objects that are not directly related to the medical instrument classification problem. However, due to the complexity of the real surgery environments, this study is limited to the classification and recognition of extracted images, and studies using real surgery videos are planned in the future. The following sections in this paper will discuss the training and testing datasets used in this study, the architectures of Faster R-CNN and SSD, and the experimental results of these architectures on our dataset, which will be followed by conclusions including directions for future studies.

2. Methods

2.1. Dataset

The dataset in this study was generated from extracted frames of around 950+ video clips provided by the Department of Neurosurgery of the Chang Gung Memorial Hospital located in Taoyuan, Taiwan [15]. They were pre-processed so that all personal information were removed and only surgical instruments were visible. Then, individual frames containing instruments were extracted from the clips with the specified requirements that each instrument must be entirely visible and that no instrument overlaps with other instruments. We ended up with 10,500+ samples of fourteen types of instruments used in the clips provided by the hospital, which are then manually labeled. The fourteen types of instruments include the following: the Clamp Pliers, the Diagonal Pliers, the regular Forceps, the Gunshaped Forceps, the modified Gunshaped Forceps (labeled as Gunshaped Forceps+), the Long Scapel, the Medium Scalpel, the Short Scapel, the Mosquito Clamp, the Mosquito Scissors, the regular Scissors, the Needle Holder, the Steel Push, and the Towel Clamp. These are the types of surgical instruments used in most of the surgeries in the neurosurgery department in the local hospital. Examples of some of the 14 instruments are shown below in Figures 1 and 2, respectively.

Figure 1. Examples of the instruments to be recognized in this study.

Figure 2. Examples of each of the instruments to be recognized, in no particular order.

The training sets and testing sets are randomly generated from these samples, so that 70% of the total samples are used for training and 30% for testing, and since this is just an exploratory study, no validation set will be used. The numbers of test data samples for each instrument are as follows in Table 1.

Table 1. The number of test samples for each type of the 14 surgical instruments.

Surgical Instrument	Number of Test Sample
Clamp Pliers	273
Diagonal Pliers	258
Forceps (General)	289
Gunshaped Forceps	259
Gunshaped Forceps+ (Modifed)	176
Long Scapel	259
Medium Scapel	120
Short Scapel	251
Mosquito Clamp	259
Mosquito Scissors	150
Scissors (General)	279
Needle Holder	280
Steel Push	150
Towel Clamp	182

2.2. The Faster R-CNN Network

Faster R-CNN has been reported to be more accurate, about 1–2% better in terms of mean average precision, in classification while operating closer to real time, about 6 frames per second, in object detection compared with other architectures presented in the paper. It is because the problem of generating a high number of possible regions of interest is reduced by the addition of a region proposal network that can more accurately identify possible regions. The structure of the Faster R-CNN network is designed to combine the functions of feature extraction, regions-of-interest proposals, bounding box regression (for refinement purposes), and classification into a single deep learning neural network. Although it takes longer to reach a classification result, it tends to be more accurate than other near real-time classifiers based on the deep learning neural networks proposed before it. It also does not limit the size of input, as other classifiers had. Its structure is shown below in Figure 3.

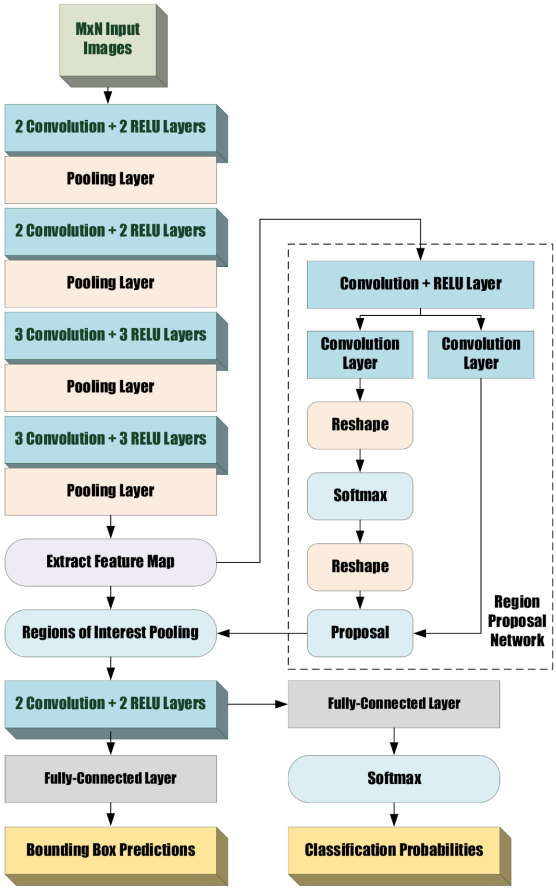

Figure 3. The structure of the Faster R-CNN network.

The convolution layers with RELU (Rectified Linear Unit) activations followed by pooling layers are used to extract interesting features from the input, which are then used to produce a feature map. The resulting feature map is sent to a region proposal network which determines whether each of the region anchors of features are positive or negative; then, it uses bounding box regression to refine the proposed regions. Then, these regions of interest deemed as positive are pooled together with the feature maps before classification. It is supposedly faster than its previous proposed convolutions-based deep learning networks without losing classification precision, as shown in various studies [16–18]. This network is chosen as one of the networks to be studied in the application of surgical instrument detection and recognition in this paper.

2.3. The Mask R-CNN Network

The Mask R-CNN Network is a modified Faster R-CNN network, and it is designed to improve bounding box detection accuracy of the Faster R-CNN network by replacing ROI-pooling operation with ROI-alignment operation. The ROI-pooling operation takes the bounding boxes found within the feature maps and overlays them to the original input images by scaling using integer values. This is faster than the ROI-alignment operation, which seeks to align the bounding boxes to the objects within the original images by non-integer scaling. In addition, in order to refine and generate the final mask, the FPN (Feature Pyramid Network) is incorporated. The claim is that this operation would result in more accurate regions of interest that would contain more useful information to the final classifier stage, and so results should be more accurate than the Faster R-CNN network. However, the alignment operation with FPN does add processing overhead when compared to the pooling operation of the Faster R-CNN network, and so the time to reach classification results would be slightly slower than the Faster R-CNN, although it is still considered as near real time. Its basic structure is shown below in Figure 4.

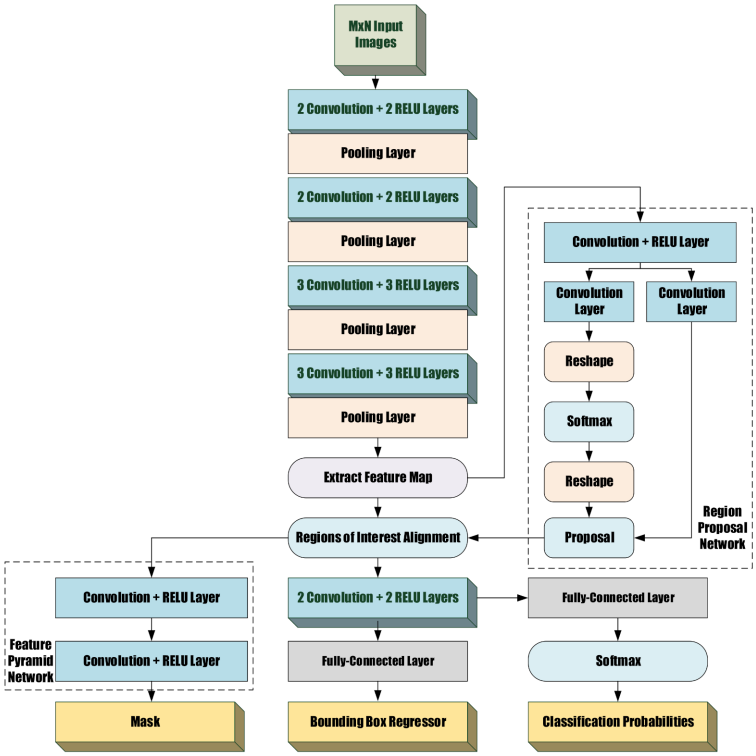

Figure 4. The structure of the Mask R-CNN network.

As can be seen in the figure above, the ROI-pooling stage of the Faster R-CNN network is replaced by an ROI-alignment stage, the output which is also used to generate a mask. This modification is supposed to achieve better detections of objects and thus better classification results.

2.4. The Single-Show Multi-Box Detector (SSD) Network

The SSD network was originally proposed to achieve a higher rate of classification than the other networks proposed before it. In terms of processing videos, it could achieve an FPS (frames per seconds) processing rate close to real time. It achieves this by using multiple convolution layers with different resolutions in order to extract interesting features at different resolutions, which will be used to generate anchors of bounding boxes at different resolutions at slightly different offsets, which is followed by non-maximum suppression before generating the final results. Its structure is shown as follows in Figure 5.

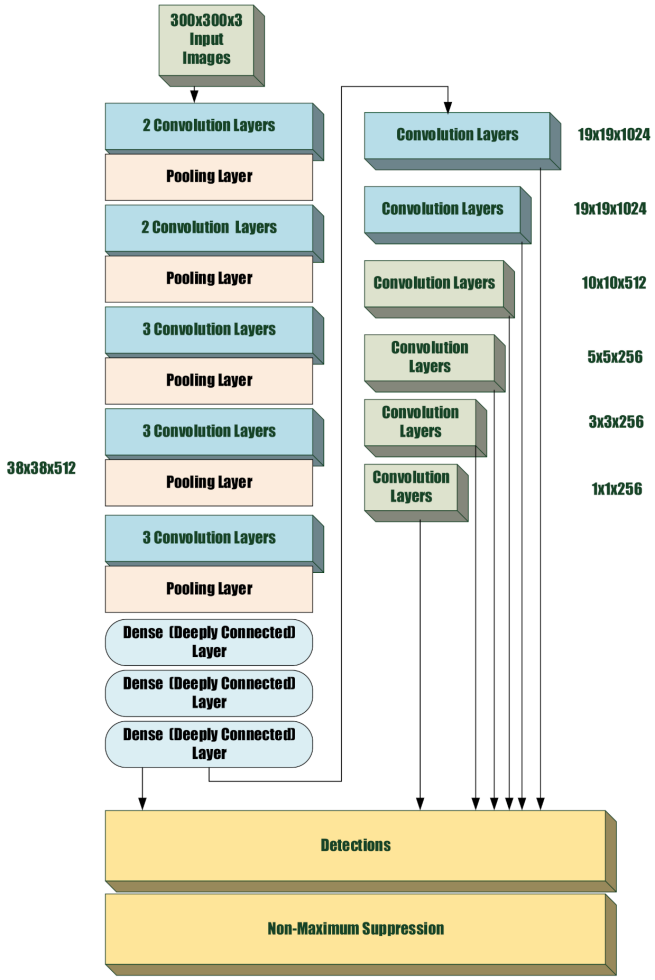

Figure 5. The structure of the SSD network.

Various studies in the literature [19,20] show that performance SSD in other fields of applications exhibits similar accuracy to Faster R-CNN, and so, it was chosen as one of the base networks to study in the application of surgical instrument detection and recognition. Since there is a difference in the design of the base architecture relative to the Faster R-CNN network, it can used to compare and determine if it is the better architecture to pursue.

2.5. Comparison of the Networks

The structure of SSD is clearly a one-shot structure, where the data and the anchors of bounding boxes are passed once through the layers to directly reach the detector, while the Faster R-CNN and Mask R-CNN, because of the use of Region Proposal Network, are actually two-shot networks. This implies that the SSD deep learning network should be able to reach a prediction for each class faster than the Faster R-CNN network and achieve higher values in raw processing rate using surgery videos, in which case the SSD network would be able to process more frames per second than the Faster R-CNN network. However, the accuracies of classification, in the application of surgical instrument detection and recognition, would still require the study of experimental results. The following section would present the experimental results of classification using these networks.

3. Experimental Results

This study is designed to test the accuracies of the unmodified Faster R-CNN, Mask R-CNN, and the Single Shot Multi-Box Detector (SSD) on the detection and identification of surgical instruments used in the Department of Neurosurgery of the Chang Gung Memorial Hospital in Taoyuan, Taiwan. The test platform used is a PC (personal computer)

with an Intel® I7 processor with 8 G of RAM and no GPU acceleration, using the Python programming language [21]. The samples are randomly separated into training and testing datasets, according the 70–30% (training–testing) principle. Then, all three networks are trained and then tested using these datasets, which is repeated for a total of three times in order to obtain the average mean precision for each type of instrument.

However, it is proposed that the Faster R-CNN network might achieve the same or more accurate results if, instead of training from scratch, fine-tuning by transfer learning is utilized [22]. The purpose of transfer learning is to take a deep learning network pre-trained using another dataset with similar features and then fine-tune it by adding training just for the features that were not in the original dataset in order to shorten the training time for the new dataset. So, for the purpose of determining whether transfer learning will both shorten the training time as well as increase classification accuracy for this application, this study will add, in addition to training the three networks using raw data, an additional experiment. That additional experiment is training (or fine-tuning) a pre-trained Faster R-CNN network that was trained using the data from the COCO (Common Objects in COntext) object detection dataset [23], which has more than 200,000 labeled common objects (no surgical instrument), as a possible solution to speed up training and hopefully achieve similar or better performance. The experimental results will be presented along with the other results. An example of a correct identification by all the networks in this study is shown below in Figure 6.

Figure 6. A pair of scissors (general) correctly detected and identified by all networks.

As expected, the speed of identification on the SSD classifier is more than twice as fast as the Faster R-CNN classifier, and the MASK R-CNN classifier is slower than the Faster R-CNN classifier. However, the SSD network sometimes failed to detect the instrument, thus lowering its average accuracy, since it is determined by the number of times each instrument is correctly identified divided by the total test samples for that instrument. The confusion matrices for one run of the experiment are shown as an illustration in Appendix A, Table A1. From just this one run, it is computed that for the Multi-Box SSD method, the average accuracy was 98.92%, the average precision was 90.96%, and the average recall was 93.76%. For Faster-RCNN (COCO), the average accuracy was 98.74%, the average precision was 90.55%, and the average recall was 91.35%. For the Faster RCNN method, the average accuracy was 99.57%, the average precision was 97.27%, and the average recall was 97.54%. For the Mask RCNN method, the average accuracy was 99.53%, the average precision was 98.96%, and the average recall was 99.24%.

In terms of accuracies, the following table, Table 2, shows the mean average precision (MAP) values for these networks after three trials along with the 95% confidence interval, where the numbers in bold show the best result, relatively, for that particular instrument.

Table 2. Classification results of the networks trained used surgical instrument data with a 95% confidence interval.

Surgical Instrument	Faster R-CNN, MAP	Faster R-CNN (COCO), MAP	Mask R-CNN, MAP	Multi-Box SSD, MAP
clamp_pliers	99.63% ± 0.2%	96.34% ± 2.6%	**100%** ± 0%	93% ± 2.7%
diagonal_pliers	**99.74**% ± 0.2%	88.76% ± 6.9%	99.26% ± 0.2%	94.52% ± 4.2%
forceps	**99.65**% ± 0.2%	86.85% ± 0.3%	99.51% ± 0.2%	88.15% ± 1.3%
gunshaped_forceps	89.19% ± 3%	73.36% ± 5.4%	**99.21**% ± 0.3%	93.82% ± 3.9%
gunshaped_forceps+	**100.00**% ± 0%	98.86% ± 0.5%	99.5% ± 0.3%	92.87% ± 4.2%
long_scalpel	**99.23**% ± 0.3%	81.08% ± 12.6%	99.21% ± 0.3%	94.94% ± 3.7%
medium_scalpel	**100.00**% ± 0%	98.33% ± 0.7%	99.49% ± 0.3%	100.00% ± 0%
mosquito_clamp	93.82% ± 0.2%	79.54% ± 10%	**99.21**% ± 0.7%	89.91% ± 0.3%
mosquito_scissors	**99.78**% ± 0.2%	98% ± 0.3%	99.07% ± 0.2%	96.33% ± 1.1%
scissors	97.85% ± 0.6%	63.80% ± 5.3%	**100**% ± 0%	87.66% ± 11.3%
short_scalpel	**99.60**% ± 0.2%	87.25% ± 2.1%	99.14% ± 0..3%	99.20% ± 0.5%
spring_needle_holders	93.57% ± 0.2%	96.97% ± 3%	**98.01**% ± 0.5%	87.09% ± 7.1%
steel_push	99.78% ± 0.3%	100% ± 0%	**100**% ± 0%	94.79% ± 2.8%
towel_clamp	86.26% ± 3.9%	78.02% ± 6.8%	**93.18**% ± 2.1%	85.24% ± 10%

Since the sample size is small, $n = 3$, the 95% confidence interval is presented using Student's t-test with the degree of freedom (df) as 2, using sample deviation. The equation for calculating a 95%, $100(1 - \alpha)$% for $\alpha = 5$, confidence interval using Student's t-test is defined as:

$$\bar{x} \pm t_{\alpha/2}\left(\frac{\text{sample deviation}}{\sqrt{n}}\right) \quad (1)$$

where $\bar{x}$ is the sample mean, and $n = 3$. The sample deviation is calculated using the following equation:

$$\text{sample deviation} = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}}. \quad (2)$$

The 95% confidence interval, in terms of percentage, is used for our experiment. In additional, for better visualization of the data, the results have been plotted in Figure 7.

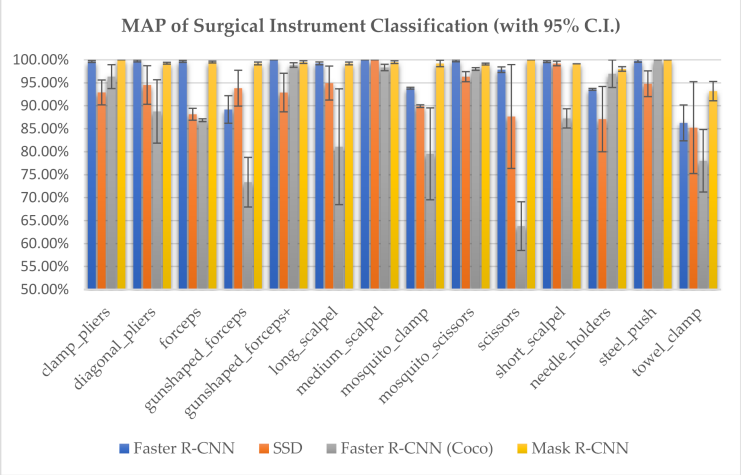

Figure 7. MAP with 95% confidence interval.

Figure 8 show examples of correctly identified instrument by the Mask RCNN network. Figure 9 shows a few examples of classified images of a clamp plier and a mosquito clamp that were correctly identified by the Mask R-CNN classifier but were mis-identified, at least once, by the original Faster R-CNN (not transfer learned) classifier.

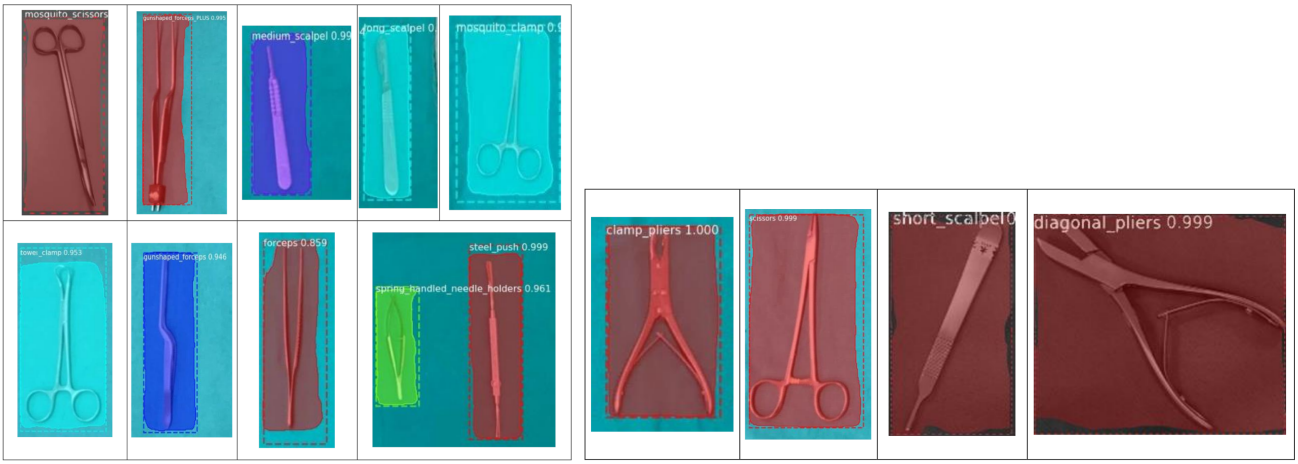

Figure 8. Some examples of correct results by the trained Mask RCNN network.

Figure 9. Examples of instruments correctly identified by Mask R-CNN but mis-identified at least once by Faster R-CNN.

The performances of the trained Faster R-CNN and the Mask R-CNN are comparable for most of the instruments, as expected. Although these classifiers took longer to detect and recognize the surgical instrument(s), they are more accurate than the SSD and the transfer learned Faster R-CNN classifers in most of the categories. The accuracies of classification for the towel clamp are not high for all networks, due to mis-identifying it as one of the other types of clamp, such as the mosquito clamp. In terms of the average MAP for all the instruments, the Mask R-CNN classifier has the highest value of 98.94%, which is followed by 96.97% achieved by the Faster R-CNN classifier without transfer learning, while the SSD classifier achieved an average MAP value of 92.61%, and the Faster R-CNN classifier using transfer learning from the COCO dataset had only reached an MAP of 87.68%. This implied that the method of transfer learning, which seeks to reduce training time, may not be suitable for the classifiers for surgical instruments—at least classifiers trained using the COCO dataset.

4. Discussion

From the experimental data above, although the performance of the Faster R-CNN and the Mask R-CNN classifiers are almost the same in most categories, in the gunshaped forceps category, the Mask R-CNN classifier outperformed the Faster R-CNN, on average, by about 10%, which is significant. However, the architecture of the Faster-RCNN is almost the same as that of the Mask R-CNN, and their main difference in the architectures would only result in the variation of the positions of the bounding boxes, so it can be claimed that a more accurate bounding box will result in a better classification in some cases. However, observing the results shown above in Figure 9, the differences in the bounding boxes generated by the two classifiers is that the bounding boxes generated by the Mask R-CNN classifier show consistent and better fits around the instrument to be identified. On the other hand, the boxes generated by the Faster R-CNN classifier are not as consistent; i.e., they are sometimes looser and sometimes tighter than those generated by the Mask R-CNN classifier. However, it is likely that bounding boxes may not be the only difference that caused the 10% difference in the gunshaped forceps category. The positions and poses of the instrument may also contribute to the differences in classification result.

In this study, only images extracted from clips were used in the experiments rather than the clips themselves, so the frames-per-second (FPS) performances of these networks may only be estimated by the amount of time it took to detect and classify each image. The single-shot method clearly outperforms the other architectures in this area; the multi-box SSD method could probably do around 16 to 20 FPS, while the other two-shot architectures could do around 5 to 6 FPS. However, since this paper considers surgical videos in which surgical instruments are not considered as fast-moving objects, 5 to 6 FPS in processing time can be considered as near real time. In addition, in examining the MAPs with confidence intervals in Table 2 and Figure 7, the high ends of the single-shot method are comparable with the top performers. This observation may imply that a good design of single-shot deep learning architecture may equal the top performers in accuracy while achieving more FPS in real-time processing for actual video processing. Another limitation of this study is that the extracted images were selected in the case where each instrument is clearly separated by itself, but in the surgical video clips, there are many cases where the instruments overlay each other. The performances of top performers may not be as expected in these cases. These cases and others will have to be included in future studies.

In terms of the experiments performed in this study, several observations can also be made from the data above. First, the performances of these networks are generally good on scalpel-type instruments but not as well on clamp-type instruments. This result may be due to the possibility that different types of clamps could exhibit similar features. The second observation is that the Faster R-CNN-based networks, including Mask R-CNN, with the region proposal network, are less likely to fail to detect the presence of a surgical instrument in comparison with the single-shot SSD network. However, whether this observation is true for all single-shot networks, such as different versions of the YOLO (You Only Look Once) network [24–26], have to be investigated in a different study. The third observation is that, as expected, the single-shot SSD network is faster in reaching a result than the two-shot Faster R-CNN. If real-time processing is a critical requirement, then one-shot networks should be investigated. However, though sometimes the SSD network did fail in detecting the instruments, the SSD classifier was able to have a slightly better performance on the gunshaped forceps than the Faster R-CNN classifier but not on the modified gunshaped forceps (gunshaped forceps+). This may be the same problem as the difference between the Mask R-CNN and Faster R-CNN networks, but it is also possible that for Faster R-CNN, the shapes and poses of the gunshaped forceps in the dataset can result in less accurate classifications when simple ROI-pooling was used. This problem was not an obvious problem in the Mask R-CNN classifier, which uses ROI-alignment, resulting in bounding boxes better fitted to the instrument to be identified.

5. Conclusions

The purpose of this study is to investigate the detection and classification performances of three near-real-time deep learning networks on static image frames containing 14 different types of surgical instruments in order to find the base architecture, the Faster R-CNN, to extend for future studies. This experiment not only found a good base architecture for time vs. accuracy tradeoff for the purpose of classifying surgical instruments used in the local hospital but also determined that by examining the MAP of each class of surgical instrument, the clamp-type instruments may exhibit similar features, thus making misclassification possible. So, a future study may include more convolution layers in order to extract more details from the clamp-type instruments. In addition, by examining the confidence intervals, the single-shot architectures may also be considered as possible candidates in future studies. In future studies, more experiments with different video clips are planned. In this study, the classifiers chosen for comparison are the Faster R-CNN, the Mask R-CNN, and the Multi-Box SSD (SSD) classifiers. According to the experimental results, the Faster R-CNN-based classifiers, though slower, achieve better accuracy by being able to detect the presence of instruments and correctly classify them most of the time. As shown in Table A3, the numbers in bold that show the best results, in terms of MAP, for that particular instrument are equally distributed between the columns of Faster R-CNN and the Mask R-CNN, from 93.18% for the towel clamp to 100% for the clamp pliers. The misclassifications of clamp-type instruments may be solved by a more precise location of the bounding boxes around the instruments, as can be seen in the performance of the Mask R-CNN classifier. However, if the classifiers only need to keep count of the numbers of instruments before, during, and after the surgery, then the Faster R-CNN detector can be of great assistance in the surgery room, since it can detect all instruments as well as the Mask R-CNN detector. Future studies will include testing other deep-learning network classifiers, choosing the classifier(s) with the best performance and extending it, and actual testing of the resultant classifier(s) on real surgery videos. The current conclusion that can be claimed in this study is that the Faster R-CNN architecture could be the base architecture to extend for this application, and it may be possible to locate a better method than the ROI-pooling operation or the ROI-alignment in order to achieve a faster classification during surgery.

Author Contributions: Conceptualization, J.-D.L. and J.-C.C.; methodology, J.-D.L. and J.-C.C.; software, J.-D.L. and Y.-T.H.; validation, J.-D.L., J.-C.C. and C.-T.W.; formal analysis, J.-D.L. and J.-C.C.; investigation, J.-D.L., J.-C.C. and C.-T.W.; resources, C.-T.W.; writing—original draft preparation, J.-D.L. and J.-C.C.; writing—review and editing, J.-D.L. and J.-C.C.; project administration, J.-D.L. and J.-C.C.; funding acquisition, J.-D.L. and J.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly supported by Ministry of Science and Technology (MOST), Taiwan, Republic of China, under Grants MOST109-2221-E-182-048 and MOST110-2221-E-182-035.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is derived from video clips provided by the Department of Neurosurgery, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan. It is currently not authorized by the hospital for release to the public domain.

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

Appendix A

Table A1. Confusion matrices for one run of the experiment using Multi-box SSD.

Multi-box SSD	clamp pliers	diagonal pliers	forceps	gunshaped forceps	gunshaped forceps+	long scalpel	medium scalpel	mosquito clamp	mosquito scissors	scissors	short scalpel	spring needle holder	steel push	towel clamp
clamp _pliers	271	4	0	11	0	0	0	0	0	0	0	0	0	0
diagonal _pliers	0	229	0	0	0	0	0	0	0	0	0	0	0	0
forceps	0	0	253	1	2	1	0	0	0	1	0	0	0	0
gunshaped _forceps	0	0	0	230	0	2	0	0	0	0	0	0	0	0
gunshaped _forceps+	0	0	0	4	172	0	0	0	0	0	1	0	0	0
long_ scalpel	0	0	0	1	0	234	0	3	0	0	1	0	0	0
medium_ scalpel	0	0	1	1	0	17	120	0	0	0	1	3	0	0
mosquito_ clamp	0	1	0	0	0	0	0	232	0	4	0	0	0	21
mosquito_ scissors	0	0	1	0	0	2	0	19	149	11	0	0	0	0
scissors	0	0	0	0	0	0	0	0	1	216	0	0	0	4
short _scalpel	0	0	0	0	0	3	0	2	0	0	248	0	0	0
spring needle holder	0	0	2	0	0	0	0	0	0	0	0	269	0	5
steel_ push	0	0	0	0	0	0	0	3	0	0	0	0	147	0
towel_ clamp	0	6	0	0	0	0	0	0	0	47	0	6	0	127
no_label (undetected)	2	18	32	11	2	0	0	0	0	0	0	2	3	25

Table A2. Confusion matrices for one run of the experiment using transfer-learnt (COCO) Faster RCNN.

Faster RCNN (COCO)	clamp pliers	diagonal pliers	forceps	gunshaped forceps	gunshaped forceps+	long scalpel	medium scalpel	mosquito clamp	mosquito scissors	scissors	short scalpel	spring needle holder	steel push	towel clamp
clamp _pliers	272	0	0	0	0	0	0	0	0	0	0	0	0	0
diagonal _pliers	1	251	0	0	0	0	0	0	0	0	0	1	0	0
forceps	0	0	252	1	1	2	0	0	0	0	0	5	0	0
gunshaped _forceps	0	0	13	207	0	0	0	0	0	0	0	0	0	0
gunshaped _forceps+	0	0	1	36	174	0	0	0	3	0	0	0	0	0
long_ scalpel	0	0	0	1	0	250	1	0	0	0	5	0	0	0
medium_ scalpel	0	0	8	1	0	4	118	0	0	0	10	2	0	0
mosquito_ clamp	0	0	0	0	0	0	0	237	0	27	0	0	0	5
mosquito_ scissors	0	0	0	2	0	0	0	0	147	19	0	0	0	0
scissors	0	0	0	0	0	0	0	10	0	182	0	0	0	13
short _scalpel	0	0	0	0	0	0	0	0	0	0	226	0	0	0
spring needle holder	0	0	3	2	0	0	0	0	0	0	0	262	0	5
steel_ push	0	0	8	0	0	0	1	6	0	0	8	0	150	0
towel_ clamp	0	5	0	4	0	0	0	4	0	51	0	4	0	157
no_label (undetected)	0	2	5	5	1	3	0	2	0	0	2	6	0	2

Table A3. Confusion matrices for one run of the experiment using Faster RCNN.

Faster RCNN	clamp pliers	diagonal pliers	forceps	gunshaped forceps	gunshaped forceps+	long scalpel	medium scalpel	mosquito clamp	mosquito scissors	scissors	short scalpel	spring needle holder	steel push	towel clamp
clamp _pliers	271	2	0	0	0	0	0	0	0	0	0	0	0	0
diagonal _pliers	2	256	0	0	0	0	0	0	0	0	0	1	0	0
forceps	0	0	287	10	1	1	0	0	0	0	0	5	0	0
gunshaped _forceps	0	0	0	241	0	2	0	0	0	0	0	0	0	0
gunshaped _forceps+	0	0	0	4	172	0	0	0	0	0	0	0	0	0
long_ scalpel	0	0	0	0	0	254	0	0	0	0	0	0	0	0
medium_ scalpel	0	0	3	0	0	0	120	0	0	0	0	1	0	0
mosquito_ clamp	0	0	0	0	0	0	0	243	0	0	0	0	0	6
mosquito_ scissors	0	0	0	0	0	0	0	0	150	2	0	0	0	0
scissors	0	0	0	0	0	0	0	12	0	274	0	0	0	13
short _scalpel	0	0	0	0	0	0	0	0	0	0	250	0	0	0
spring needle holder	0	0	0	2	0	0	0	0	0	0	0	271	0	5
steel_ push	0	0	0	0	0	0	0	0	0	0	0	0	150	0
towel_ clamp	0	0	0	0	0	0	0	4	0	1	1	0	0	157
no_label (undetected)	0	0	0	2	2	0	0	0	0	2	0	2	0	1

Table A4. Confusion matrices for one run of the experiment using Mask RCNN.

Masked RCNN	clamp pliers	diagonal pliers	forceps	gunshaped forceps	gunshaped forceps+	long scalpel	medium scalpel	mosquito clamp	mosquito scissors	scissors	short scalpel	spring needle holder	steel push	towel clamp
clamp _pliers	273	0	0	0	0	0	0	0	0	0	0	0	0	0
diagonal _pliers	0	258	0	0	0	0	0	0	0	0	0	0	0	0
forceps	0	0	289	0	0	1	0	0	0	0	0	1	0	0
gunshaped _forceps	0	0	0	256	0	0	0	0	0	0	0	0	0	0
gunshaped _forceps+	0	0	0	2	176	0	0	0	0	0	0	0	0	0
long_ scalpel	0	0	0	0	0	257	0	0	0	0	1	0	0	0
medium_ scalpel	0	0	0	0	0	0	120	0	0	0	0	1	0	0
mosquito_ clamp	0	0	0	0	0	0	0	258	0	3	0	0	0	0
mosquito_ scissors	0	0	0	0	0	0	0	0	150	0	0	0	0	0
scissors	0	0	0	1	0	0	0	1	0	273	0	0	0	2
short _scalpel	0	0	0	0	0	0	0	0	0	0	250	0	0	0
spring needle holder	0	0	0	0	0	0	0	0	0	0	0	274	0	5
steel_ push	0	0	0	0	0	0	0	0	0	0	0	0	150	0
towel_ clamp	0	0	0	0	0	0	0	4	0	1	1	0	0	157
no_label (undetected)	0	0	0	2	2	0	0	0	0	2	0	2	0	1

References

1. Wubben, I.; van Manen, J.G.; van den Akker, B.J.; Vaartjes, S.R.; van Harten, W.H. Equipment-related incidents in the operating room: An analysis of occurrence, underlying causes and consequences for the clinical process. *J. Qual. Saf. Health Care* **2010**, *6*, 5. [CrossRef] [PubMed]
2. Guédon, A.C.P.; Wauben, L.S.G.L.; Van Der Eijk, A.C.; Vernooij, A.S.N.; Meeuwsen, F.C.; Van Der Elst, M.; Hoeijmans, V.; Dankelman, J.; Dobbelsteen, J.J.V.D. Where are my instruments? Hazards in delivery of surgical instruments. *Surg. Endosc.* **2016**, *30*, 2728–2735. [CrossRef] [PubMed]
3. Hosaka, R.; Noji, R. Automatic identification for surgical instruments using UHF band passive RFID. In Proceedings of the VI Latin American Congress on Biomedical Engineering CLAIB 2014, Paraná, Argentina, 29–31 October 2014; Springer Science and Business Media LLC.: Singapore, 2017; Volume 65, pp. 1061–1064.
4. Lin, Q.; Cai, K.; Yang, R.; Chen, H.; Wang, Z.; Zhou, J. Development and Validation of a Near-Infrared Optical System for Tracking Surgical Instruments. *J. Med. Syst.* **2016**, *40*, 1–14. [CrossRef] [PubMed]
5. Zhou, T.; Wachs, J.P. Needle in a haystack: Interactive surgical instrument recognition through perception and manipulation. *Robot. Auton. Syst.* **2017**, *97*, 182–192. [CrossRef]
6. Cai, T.; Zhao, Z. Convolutional neural network-based surgical instrument detection. *Technol. Health Care* **2020**, *28*, 81–88. [CrossRef] [PubMed]
7. EndoVis Grand Challenge. Available online: https://endovis.grand-challenge.org/ (accessed on 30 May 2021).
8. 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. [CrossRef] [PubMed]
9. 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] [PubMed]
10. Yu, L.; Wang, P.; Yan, Y.; Xia, Y.; Cao, W. MASSD: Multi-scale attention single shot detector for surgical instruments. *Comput. Biol. Med.* **2020**, *123*, 103867. [CrossRef] [PubMed]
11. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; IEEE, Manhattan, NY, USA. pp. 2980–2988.
12. Sanchez, S.A.; Romero, H.J.; Morales, A.D. A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework. *IOP Conf. Series: Mater. Sci. Eng.* **2020**, *844*, 15. [CrossRef]
13. Hui, J. Object Detection: Speed and Accuracy Comparison. Available online: https://jonathan-hui.medium.com/object-detection-speed-and-accuracy-comparison-faster-r-cnn-r-fcn-ssd-and-yolo-5425656ae359 (accessed on 15 September 2020).
14. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *Lect. Notes Comput. Sci.* **2016**, *9905*, 21–37.
15. Chang Gung Memorial Hospital at Linkou, Taiwan. Available online: https://www1.cgmh.org.tw/branch/lnk/2016/en/ (accessed on 8 March 2020).
16. Wan, S.; Goudos, S. Faster R-CNN for multi-class fruit detection using a robotic vision system. *Comput. Netw.* **2020**, *168*, 107036. [CrossRef]
17. Zhang, C.; Xu, X.; Tu, D. Face Detection Using Improved Faster RCNN. *arXiv* **2018**, arXiv:1802.02142.
18. Ren, Y.; Zhu, C.; Xiao, S. Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures. *Math. Probl. Eng.* **2018**, *2018*, 3598316. [CrossRef]
19. Nguyen, A.Q.; Nguyen, H.T.; Tran, V.C.; Pham, H.X.; Pestana, J. A Visual Real-time Fire Detection using Single Shot MultiBox Detector for UAV-based Fire Surveillance. In Proceedings of the 2020 IEEE Eighth International Conference on Communications and Electronics (ICCE), Phu Quoc Island, Vietnam, 13–15 January 2021; pp. 338–343.
20. Kanimozhi, S.; Gayathri, G.; Mala, T. Multiple Real-time object identification using Single shot Multi-Box detection. In Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 21–23 February 2019; pp. 1–5.
21. Ge, W.; Yu, Y. Borrowing Treasures from the Wealthy: Deep Transfer Learning through Selective Joint Fine-Tuning. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, HI, USA, 21–26 July 2017; pp. 10–19.
22. Common Objects in Context (COCO). Available online: https://cocodataset.org/ (accessed on 15 April 2021).
23. Python. Available online: https://www.python.org/ (accessed on 8 February 2021).
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, CA, USA, 26 June–1 July 2016; pp. 779–788.
25. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Hawaii, HI, USA, 21–26 July 2017; pp. 6517–6525.
26. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.