applied sciences

IMPACT FACTOR 2.5

CITESCORE 5.5

Article

Surgical Instrument Recognition Based on Improved YOLOv5

Kaile Jiang, Shuwan Pan, Luxuan Yang, Jie Yu, Yuanda Lin and Huaiqian Wang

MDPI

https://doi.org/10.3390/app132111709

applied sciences

MDPI

Article

Surgical Instrument Recognition Based on Improved YOLOv5

Kaile Jiang [1], Shuwan Pan [2,*], Luxuan Yang [1], Jie Yu [2], Yuanda Lin [2] and Huaiqian Wang [1,2]

[1] College of Information Science and Engineering, Huaqiao University, Xiamen 361021, China; 23013082008@stu.hqu.edu.cn (K.J.); 23013082014@stu.hqu.edu.cn (L.Y.); hqwang@hqu.edu.cn (H.W.)
[2] College of Engineering, Huaqiao University, Quanzhou 362021, China; 22014084023@stu.hqu.edu.cn (J.Y.); 23014084008@stu.hqu.edu.cn (Y.L.)
* Correspondence: shuwanpan@hqu.edu.cn; Tel.: +86-0595-2269-4169

Abstract: Recognition of surgical instruments is a key part of the post-operative check and inspection of surgical instrument packaging. However, manual inventorying is prone to counting errors. The achievement of automated surgical instrument identification holds the potential to significantly mitigate the occurrence of medical accidents and reduce labor costs. In this paper, an improved You Only Look Once version 5 (YOLOv5) algorithm is proposed for the recognition of surgical instruments. Firstly, the squeeze-and-excitation (SE) attention module is added to the backbone to improve the feature extraction. Secondly, the loss function of YOLOv5 is improved with more global parameters to increase the convergence rate of the loss curve. Finally, an efficient convolution algorithm is added to the C3 module in the head to reduce computational complexity and memory usage. The experimental results show that our algorithm outperforms the original YOLOv5 with improvements observed across various metrics: mean average precision 50–95 (mAP50-95) achieved 88.7%, which improved by 1.8%, and computational requirements reduced by 39%. This study, with a simple but effective method, is expected to be a guide for automatically detecting, classifying, and sorting surgical instruments.

Keywords: YOLOv5; surgical instruments recognition; surgical instruments dataset; loss function; attention mechanism; deep-learning

Citation: Jiang, K.; Pan, S.; Yang, L.; Yu, J.; Lin, Y.; Wang, H. Surgical Instrument Recognition Based on Improved YOLOv5. *Appl. Sci.* **2023**, *13*, 11709. https://doi.org/10.3390/app132111709

Academic Editor: Dosik Hwang

Received: 24 September 2023
Revised: 14 October 2023
Accepted: 23 October 2023
Published: 26 October 2023

1. Introduction

Manual inventorying of surgical instruments is a critical step after completing surgery. After sterilization or post-operative, surgical instruments need to be carefully counted and verified by medical staff. However, manual inventorying is prone to counting errors. Manual inventorying can also be time-consuming, especially in complex surgeries where a large number of instruments are used. Therefore, there is a great practical need for fast and accurate surgical instrument detection techniques. With the continuous advancement of medical technology, the types of surgical instruments are being expanded and updated, which increases the difficulty of counting surgical instruments. Thus, automatic instrument detection and identification has become an important research area for improving surgical safety and efficiency, as well as reducing the workload of medical staff [1].

In recent years, object detection technology combined with deep learning has developed rapidly. R-CNN (regions with CNN features), proposed by Ross Girshick et al. in 2013 [2], is one of the first object detection models based on deep learning. The R-CNN architecture consists of three main components: region proposal generation, feature extraction using a CNN, and classification using SVMs [3]. However, it is computationally expensive due to the need to process each region proposal separately, making it unsuitable for real-time applications. Faster R-CNN is a two-stage object detection model that improves on its predecessors, R-CNN and Fast R-CNN, by using a region proposal network (RPN) with the CNN model [4,5]. In the ILSVRC and COCO 2015 competitions, Faster R-CNN and RPN are the basis of the first-place winning entries in several tracks. Wei Liu et al. proposed

the Single Shot MultiBox Detector (SSD) in 2015. It is designed to detect multiple object categories in a single forward pass through the network, making it suitable for real-time applications. In addition, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of different sizes [6,7]. The original FPN architecture was introduced by Tsung-Yi Lin et al. in 2016. FPNs combine features from different layers of a convolutional network to improve detection performance [8,9]. The FPN architecture is independent of the backbone convolutional architectures, making it a generic solution for building feature pyramids within deep convolutional networks.

Deep learning has also made remarkable progress in the field of surgical instrument detection. Sabrina Kletz et al. investigated the segmentation and detection of surgical instruments in laparoscopic gynecology videos based on Mask R-CNN [10]. Sheng Wang et al. proposed a multi-label classification method to detect the presence of surgical instruments in laparoscopic videos. They experimented with two methods, one using VGGNet and the other using GoogleNet. Then, they combined both models, and the ensemble results produced better results than the single model [11]. Hiroyuki Sugimori et al. used a convolutional neural network based on You Only Look Once version 2 (YOLOv2) as the network model for training and nine video recordings of carotid endarterectomies for training and testing [12]. Jani Koskinen et al. introduced and investigated a method for simultaneous detection and processing of micro-instruments and gaze during microsurgery [13]. They trained and evaluated a fast CNN-based detector to perform tool detection from a dataset of approximately 20 surgical positions and 17 micro-instruments.

However, in the past, the application of deep learning in the medical field usually focused on the detection of the position of surgical instruments during surgery to enable machine control of surgical instruments. There has been less relevant application for the detection of surgical instruments before and after use. In view of the various defects of manual detection of surgical instruments, this article proposes the automation of surgical instrument detection, combines it with the current image recognition field of deep learning, and improves the YOLOv5 to make it suitable for surgical instrument recognition.

The main works of this paper are as follows: (a) We establish a standard surgical instrument dataset based on eight types of representative surgical instruments and extended by a random-mixed image augmentation method. (b) We optimize the backbone network combined with the attention mechanism to make it more suitable for the detection of surgical instruments. (c) We use the dynamic sparse convolution algorithm instead of ordinary convolution to reduce the computational complexity. (d) We optimize the loss function and replace the original IoU metric with Wise-IoU.

By using the proposed method, it is anticipated that this work will serve as a guide and provide solutions to the challenges of recognizing surgical instruments for automatic sorting.

2. Materials and Methods

2.1. Images Dataset and Augmentation

In this study, we utilized a blue non-woven fabric commonly found in surgical instrument inventories as the background for our dataset. We carefully selected eight representative surgical instruments for inclusion. A total of 740 original images were collected to construct our comprehensive surgical instrument dataset. Among these, 540 images depicted non-stacked surgical instruments, while the remaining 200 images showcased stacked surgical instruments across the eight instrument types. Each image was manually labeled by our team. Figure 1 provides visual examples of these labeled images.

The dataset consists of eight representative surgical instruments: hemostat, speculum, napkintong, scissors, tweezers, colposcope, attractor, and stripper. These instruments were selected to encompass distinct inter-class differences and minimal intra-class differences. Figure 2 provides examples of some of the eight surgical instruments.

(**a**) (**b**)

Figure 1. Surgical instruments image examples in the dataset: (**a**) is the example of non-stacked; and (**b**) is the example of stacked surgical instruments.

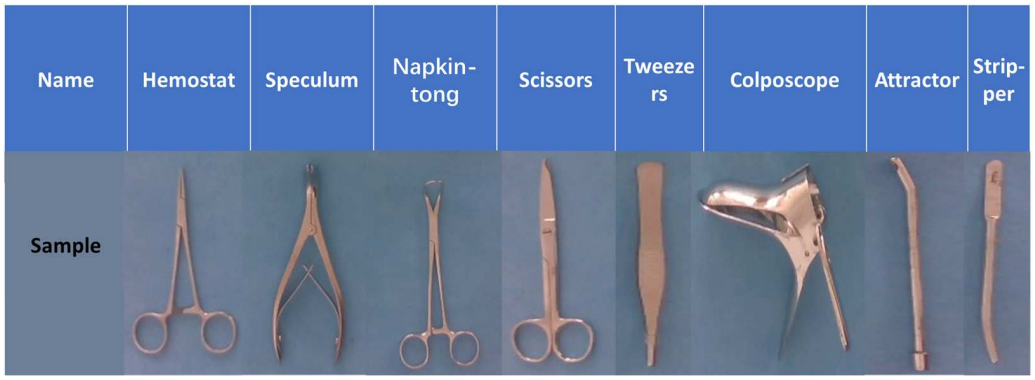
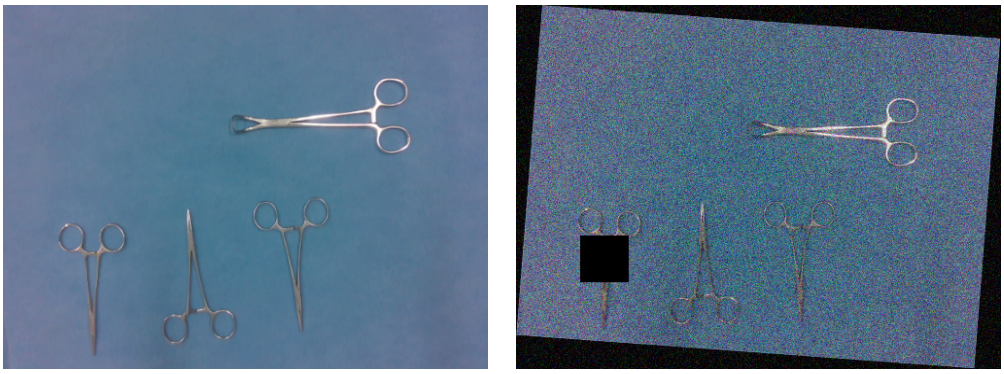


Figure 2. Examples of the eight surgical instruments.

Due to the limited number of images available in the dataset of surgical instruments, the training process of deep learning is highly at risk of overfitting. This requires a substantial amount of data and diverse regularization methods. Data augmentation is considered as one of these methods. We applied techniques of image augmentation to increase the dataset and thereby reduce the threat of overfitting in the neural network. A random-mixed image augmentation method was applied, integrating seven essential image augmentation techniques: panning, cropping, brightness adjustment, noise introduction, angle rotation, horizontal flipping, and the cutout technique. Cutout, a straightforward yet potent regularization method for convolutional neural networks, was employed during the training process. Cutout is achieved by selecting a random location within the input image and setting the pixel values within a square region centered at that location to zero (or another constant value). In the random-mixed image augmentation process, these seven augmentation methods are randomly selected and incrementally applied six times, resulting in six augmented images. Figure 3 showcases an example of the results produced by the random-mixed image augmentation method.

After applying image augmentation, our dataset expanded to over 20,000 samples across the eight instrument types. We randomly generated training and testing sets from these samples, allocating 90% for training and 10% for testing. Additionally, a separate validation set was collected. Table 1 showcases the number of data samples for each instrument.



(**a**) (**b**)

Figure 3. An example of the random-mixed image augmentation method results: (**a**) the original image; and (**b**) the augmented image.

Table 1. The number of surgical instrument samples.

Name	The Original Dataset	The Augmentation Dataset	Total
Hemostat	512	3072	3854
Speculum	389	2334	2723
Napkintong	375	2250	2625
Scissors	336	2016	2352
Tweezers	338	2028	2366
Colposcope	309	1854	2163
Attractor	330	1980	2310
Stripper	388	2328	2716

2.2. YOLOv5 and Improve Methods

YOLO is currently a commonly used target detection neural network [14–16]. YOLOv5 is an object detection model that is designed for real-time object detection and is known for its speed and accuracy [17]. Regarding its applications, YOLOv5 has garnered extensive adoption across diverse domains. It has proven effective in real-time vehicle detection [18] and object detection in maritime datasets [19]. The model's remarkable capability to deliver rapid inference speeds renders it highly suitable for real-time scenarios requiring swift and precise object detection. The architecture of YOLOv5 consists of several key components: The backbone is responsible for extracting features from the input image. YOLOv5 utilizes CSPNet (Cross Stage Hierarchical Networks) as the backbone, which is an efficient and scalable architecture for feature extraction. The neck connects the backbone to the detection head and is responsible for aggregating and refining the features extracted by the backbone. YOLOv5 employs PANet (Path Aggregation Network) as the neck, which helps in improving the information flow between layers. The detection head is responsible for predicting the object bounding boxes, class probabilities, and objectness scores. YOLOv5 utilizes anchor boxes to predict the bounding box coordinates and applies a sigmoid activation function to the output for better accuracy. The architecture of the improved YOLOv5 is shown in Figure 4.

2.2.1. Squeeze-and-Excitation Attention Mechanism

The attention mechanism in deep learning serves as an approach to emulate cognitive attention within artificial neural networks. By augmenting certain parts of the input data and reducing others, the network can concentrate more effectively on the critical aspects of the data. The significance of different parts of the data is contingent upon the context, which is established through gradient descent.20.

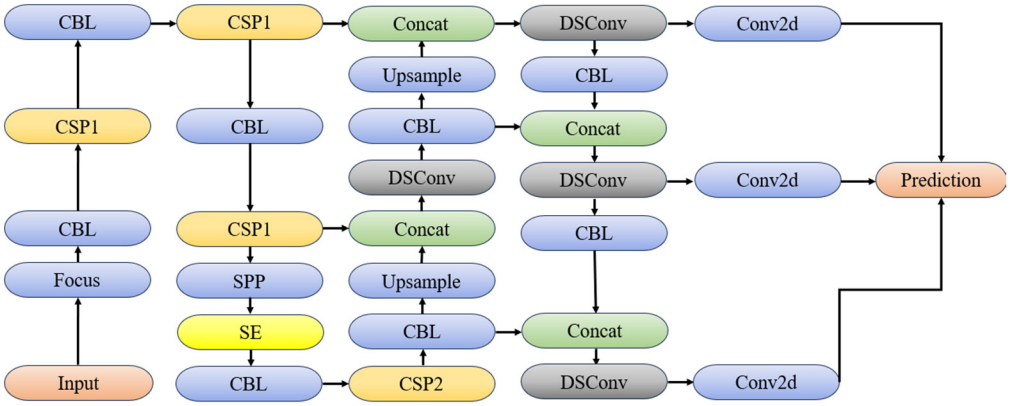


Figure 4. The architecture of the improved YOLOv5.

Squeeze-and-excitation (SE) networks are a type of channel attention mechanism that aims to improve the networks by selectively emphasizing informative features and suppressing less useful ones [20]. The SE block consists of three main operations: (a) Squeeze: This operation aims to extract global information from each channel of the feature map. It reduces the spatial dimensions of the feature map by applying global average pooling (GAP), which calculates the average value of each channel. (b) Excitation: The squeezed tensor is then passed through a fully connected multi-layer perceptron (MLP) with a bottleneck structure. The MLP is used to generate weights for adaptively scaling each channel of the feature map. It consists of two layers: the first layer reduces the number of features by a reduction factor, and the second layer restores the original number of channels. (c) Scaling: The output of the excitation operation is passed through a sigmoid activation function, which maps the tensor values to a range between 0 and 1. An element-wise multiplication is then performed between the output of the sigmoid activation function and the input feature map. This scales the feature map channels according to their importance. The structure of the SE module is shown in Figure 5.

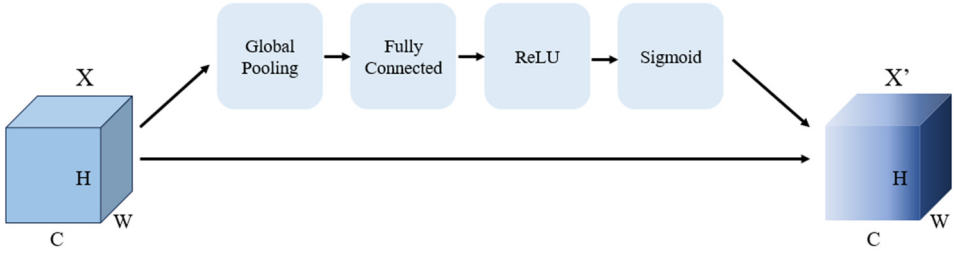


Figure 5. The SE module.

The SE module operates as follows: First, the tensor X is transformed into a $1 \times 1 \times C$ tensor Z through global average pooling. Then, after passing through a fully connected layer, a non-linear transformation, and a sigmoid activation function, the tensor is normalized to a set of real numbers ranging from 0 to 1. These real numbers represent the importance of each channel, with 1 indicating high importance and 0 indicating unimportance. Finally, the feature map is multiplied to obtain the output X'.

2.2.2. IoU, GIoU, and Wise-IoU

Intersection over union (IoU) is a metric commonly used in the context of object detection and segmentation tasks in neural networks [21]. It serves as a measure of the overlap between two bounding boxes or regions, providing an evaluation of a model's ability to predict object locations within an image accurately. IoU is calculated as the

ratio of the intersection area between two bounding boxes (or regions) to their union area. Mathematically, it can be expressed as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Then, IoU loss can be expressed as:

$$LossIoU = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

IoU values range from 0 to 1, where 0 indicates no overlap between the bounding boxes, and 1 represents a perfect match. Higher IoU values indicate better performance of the model in localizing objects. In practical applications, a threshold is often set to determine whether a predicted bounding box is classified as a true positive or a false positive. For instance, if the IoU threshold is set to 0.5, any predicted bounding box with an IoU of 0.5 or higher is considered a true positive relative to the ground truth bounding box. Conversely, predicted bounding boxes with lower IoU values are classified as false positives.

The original YOLOv5 incorporates the generalized intersection over union (GIoU) metric, which was introduced by Rezatofighi et al. in 2019 [22]. GIoU aims to overcome the limitations of IoU, especially when predicted and ground truth bounding boxes are non-overlapping. It is defined as the difference between the IoU and a shape-aware penalty term that considers the geometry of the bounding boxes. The penalty term is calculated based on the smallest enclosing bounding box that contains both the predicted and ground truth bounding boxes, called the "smallest enclosing box", denoted as C. Then, it can be expressed as:

$$GIoU = \frac{|A \cap B|}{|A \cup B|} - \frac{|C \backslash (A \cup B)|}{|C|} \tag{3}$$

GIoU values range from -1 to 1, with -1 indicating the poorest conceivable match and 1 representing a flawless match. GIoU is an advancement over the standard IoU measurement, as it takes into account the geometry of bounding boxes and offers superior gradient data for the neural network training of object detection tasks.

Nonetheless, GIoU also presents several issues, such as its sensitivity to object size scaling. This implies that the model performance may differ depending on the dimensions of the recognized objects. Sometimes, the model may find it challenging to identify smaller items or may create less precise boxing for more significant entities. Furthermore, the optimization process can cause GIoU loss function to be trapped in local minima. Consequently, this can induce suboptimal model performance, as the model may not attain the most satisfactory outcome. The Wise-IoU loss function incorporates weight factors derived from GIoU to address the issues posed by GIoU [23]. The function assigns varying weights to distinct parts of the bounding box based on their significance for object detection. The Wise-IoU loss function is defined as follows:

$$WiseIoU = IoU - \frac{|\lambda * (C - (A \cup B))|}{|C|} \tag{4}$$

Among them, γ is an adjustable parameter employed to regulate the weight factor's magnitude. The Wise-IoU loss function incorporates a weight factor λ to provide a more accurate reflection of the positional relationship between predicted and real bounding boxes. The weight factor λ increases in value when the overlap between predicted and actual bounding boxes is minimal, resulting in greater sensitivity to changes to the former. Conversely, if there is a significant overlap between the predicted and actual bounding boxes, the weight factor λ assumes smaller values, leading to a gradual adjustment of the

loss value regarding the predicted bounding box. The weight factor λ is regulated by the hyperparameter γ and IoU:

$$\lambda = 1 - e^{(-\gamma * IoU)} \quad (5)$$

The Wise-IoU loss function provides several advantages compared to the IoU and GIoU loss functions. Firstly, it effectively mitigates the issue of gradient vanishing, ensuring a more stable training process. Secondly, by introducing a weight factor, the Wise-IoU loss function better captures the relative positional relationship between predicted and ground truth bounding boxes, leading to improved object detection and instance segmentation performance. Additionally, the parameter γ can be fine-tuned according to the specific task requirements, allowing optimal results.

2.2.3. Distribution Shifting Convolution

DSConv (distribution shifting convolution) represents a specific type of convolutional layer [24] that utilizes dynamic sparsity patterns in convolutional weights, which in turn significantly reduces the computational cost and memory usage of CNNs. Through the use of variable quantised kernel (VQK) and distribution shifts, DSConv divides the traditional convolution kernel into two components, leading to improved memory usage and faster speed.

The VQK stores integer values, leading to memory savings and faster operation. Kernel-based and channel-based distribution offsets maintain the original convolution's output in DSConv. Figure 6 shows the general idea of DSConv.

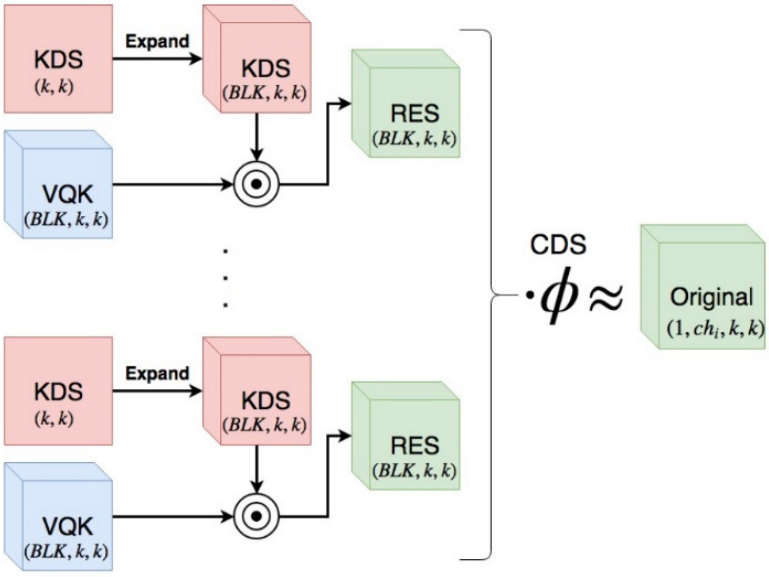


Figure 6. The general idea of DSConv [18].

3. Ablation Experiment

The ablation experiment aims to evaluate the detection and identification capabilities of the networks with various improved modules. Ablation experiments find extensive application across diverse scientific disciplines, such as biology, computer science, neuroscience, and engineering. Through the systematic removal or disabling of specific elements, researchers can observe the resultant alterations in the system's behavior, thereby inferring the relative importance of the removed component. The samples are randomly divided into training and testing datasets at a ratio of 90:10 (training: testing). Following this, the original YOLOv5 network and the six improved networks are trained and tested with these datasets. The detection indicators comprise FLOPs (floating-point operations), which gauges the calculation speed, and mAP50-95, which gauges the mean accuracy spanning the neural network's 0.5 to 0.95 IoU thresholds. Table 2 displays the ablation experiment results.

Table 2. The results of the ablation experiments.

Detection Index	The Original YOLOv5	SE	DSC	Wise-IoU	SE + DSC	SE + Wise-IoU	DSC + Wise-IoU
Hemostat's AP (%)	93.4	94.9	94.6	94.7	94.3	94	94.9
Speculum's AP (%)	94.1	93.4	92	92.9	92.1	92.5	92.3
Napkintong's AP (%)	92.2	94.3	95.1	95	95.2	95.2	95
Scissors' AP (%)	90.9	94.2	92.2	93.3	93.2	92.9	92.1
Tweezers' AP (%)	76	75.7	76.4	76.1	77.1	77.5	77.4
Colposcope's AP (%)	96.7	97	96.4	96.1	96.9	97.3	96.4
Attractor's AP (%)	75.8	77.2	79.5	79.2	79.2	79.1	78.8
Stripper's AP (%)	75.9	78.2	76.4	77.2	78.4	77.9	79.6
mAP (%)	86.9	88.1	87.8	88	88.3	88.3	88.3
FLOPs (G)	16.9	16.9	10.3	16.9	10.3	16.9	10.3

4. Experimental Results

Firstly, the improvement effect of the loss function was assessed by comparing the loss curve of YOLOv5 before and after implementation. The loss curve of a neural network represents the change in the loss function as the number of training iterations increases during the model training process. Figure 7 illustrates the comparative analysis of the algorithm's loss curve before and after the improvement. The data clearly indicate that the YOLO network with Wise-IoU achieves a faster convergence rate, as evidenced by reaching a lower loss value by the tenth batch.

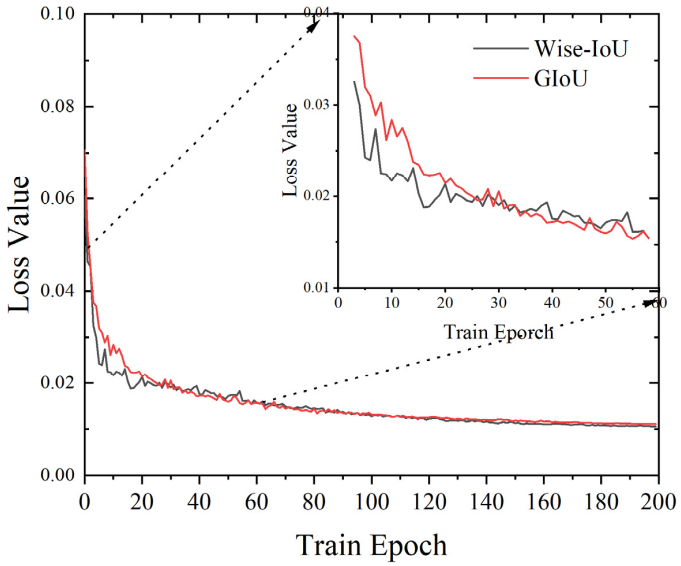


Figure 7. The loss curve comparison of Wise-IoU and GIoU.

Secondly, the accuracy comparison between the original YOLOv5 and the improved version is displayed in Table 3, using AP50-95 and mAP50-95 metrics. The improved YOLOv5 shows an increase in AP50-95 for most surgical instrument classes, except for the speculum class, compared to the original YOLOv5 and the SSD, as presented in Table 3. Despite the significantly compressed structure of the improved network model, the mAP50-95 of the algorithm in this paper is 1.8% higher than the original YOLOv5. Analyzing the SSD data reveals that the training outcomes of the SSD neural network exhibit average performance across different devices. Notably, the average precision (AP) values for most devices are lower compared to the YOLOv5 network model.

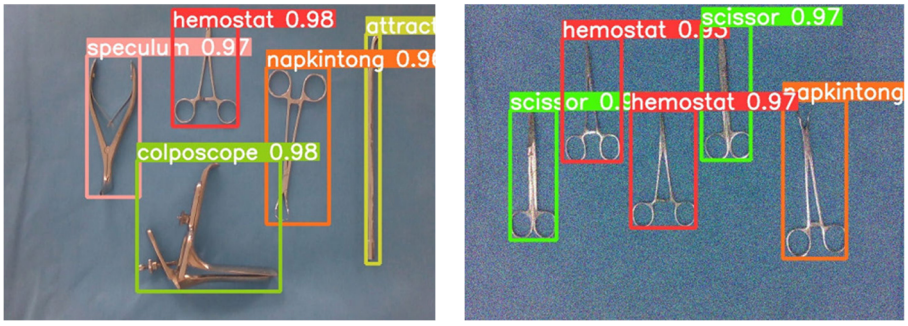
Thirdly, FLOPs was utilized as the determinant of neural network detection speed. FLOPs denotes the tally of floating-point operations that the algorithm necessitates to execute during the process. It is not related to the hardware and software specifications of the program's running environment, and it allows for a relatively fair comparison of the

detection speed between algorithms. The comparison between the floating-point operations required by the algorithm before and after enhancement and the SSD is displayed in Table 3. The FLOPs of SSD are 47.7G, which is much larger than YOLOv5. That is, the model complexity of the SSD neural network is much larger than YOLOv5. The FLOPs demand of the YOLOv5 for image detection has been considerably lowered from its original 16.9 G to 10.3 G, which reduced by 39%. This drastic reduction has expedited the neural network's detection rate.

Table 3. The comparison of AP50-95 between the improved YOLOv5 and the original YOLOv5.

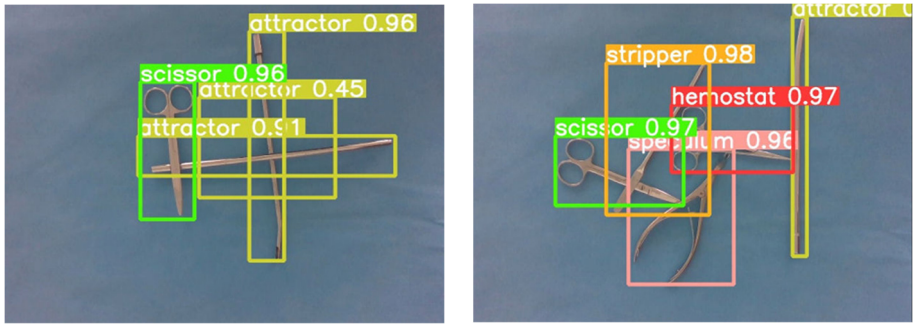
Detection Index	The Improved YOLOv5	The Original YOLOv5	SSD
Hemostat's AP (%)	94.4	93.4	71.8
Speculum's AP (%)	92.7	94.1	68.1
Napkintong's AP (%)	95.2	92.2	74.1
Scissors' AP (%)	94.6	90.9	74.6
Tweezers' AP (%)	77.1	76	67.7
Colposcope's AP (%)	97.6	96.7	71.0
Attractor's AP (%)	78.9	75.8	74.2
Stripper's AP (%)	78.8	75.9	66.0
mAP (%)	88.7	86.9	71.0
FLOPs (G)	16.9	10.3	47.7

Figure 8 presents non-stacked examples of surgical instrument images recognized by the improved network. The recognition results are mostly accurate. Figure 9 illustrates examples of stacked surgical instrument images generated by the improved network. The results indicate that while the stacked images consisting of multiple types of surgical instruments are successful, those comprising highly stacked images of multiple attractors produce false detections.



(**a**) (**b**)

Figure 8. The examples (**a**) and (**b**) of detection results of the non-stacked surgical instruments sample.



(**a**) (**b**)

Figure 9. The examples (**a**) and (**b**) of detection results of the stacked surgical instruments sample.

5. Conclusions

This work aims to propose an optimal approach for recognizing surgical instruments intended for the automatic sorting processes. In this work, we create a surgical instrument dataset that includes eight types of representative surgical instruments. We improve the YOLOv5 combined with the actual surgical instrument recognition situation. The squeeze-and-excitation attention mechanism is added to improve the feature extraction ability. The dynamic sparse convolution is used instead of the ordinary convolution for the real-time recognition of post-operative surgical instruments. The Wise-IOU loss function with the weighting factor λ is used to replace the original loss function to address the issues of slow convergence speed and easy disappearance of gradients. The results show that our algorithm increases the accuracy, reduces the FLOPs, and speeds up the convergence of the loss curve compared to the original YOLOv5. This paper presents a viable solution for the automated detection of surgical instruments post-surgery. By significantly expanding the dataset, this research is expected to bridge the gap between the proposed solution and real-world application scenarios.

Author Contributions: Conceptualization, S.P.; Methodology, S.P.; Software, S.P. and H.W.; Validation, K.J., J.Y. and Y.L.; Formal Analysis, J.Y. and Y.L.; Investigation, S.P. and K.J.; Resources, K.J. and L.Y.; Data Curation, K.J.; Writing—Original Draft Preparation, K.J., L.Y., J.Y. and Y.L; Writing—Review and Editing, S.P. and H.W.; Visualization, K.J.; Supervision, S.P.; Project Administration, S.P.; Funding Acquisition, S.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the High Level Talent Innovation and Entrepreneurship Project of Quanzhou under Grant 2021C047R; the University Industry Education Cooperation Project of Fujian Province under Grant 2022H6013; Supported by the Fundamental Research Funds for the Central Universities under Grant ZQN-1121; Collaborative Innovation Platform Project of Fuzhou-Xiamen-Quanzhou National Independent Innovation Demonstration Zone under Grant 2021FX03.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are available within the article.

Acknowledgments: This work was supported by the High Level Talent Innovation and Entrepreneurship Project of Quanzhou under Grant 2021C047R, the University Industry Education Cooperation Project of Fujian Province under Grant 2022H6013, in part by the Pilot Project of Fujian Province under Grant 2022H0017, and in part by the Quanzhou City Science and Technology Major Special Pilot Project of China under Grant 2022GZ1.

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

References

1. Xu, Y.; Tong, X.; Mao, Y.; Griffin, W.B.; Kannan, B.; DeRose, L.A. A vision-Guided Robot Manipulator for Surgical Instrument Singulation in A Cluttered Environment. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation, Hong Kong, China, 31 May–7 June 2014; pp. 3517–3523.
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
3. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef] [PubMed]
5. Singh, S.; Ahuja, U.; Kumar, M. Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment. *Multimed. Tools Appl.* **2021**, *80*, 19753–19768. [CrossRef] [PubMed]
6. 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: Cham, Switzerlan, 2016.
7. Nagrath, P.; Jain, R.; Madan, A.; Arora, R.; Kataria, P.; Hemanth, J. SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2. *Sustain. Cities Soc.* **2021**, *66*, 102692. [CrossRef] [PubMed]

8. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
9. Ghiasi, G.; Lin, T.; Le, Q.V. NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
10. Kletz, S.; Schoeffmann, K.; Benois-Pineau, J.; Husslein, H. Identifying Surgical Instruments in Laparoscopy Using Deep Learning Instance Segmentation. In Proceedings of the 2019 International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 4–6 September 2019; pp. 1–6.
11. Wang, S.; Raju, A.; Huang, J. Deep Learning Based Multi-Label Classification for Surgical Tool Presence Detection in Laparoscopic Videos. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Melbourne, VIC, Australia, 18–21 April 2017; pp. 620–623.
12. Sugimori, H.; Sugiyama, T.; Nakayama, N.; Yamashita, A.; Ogasawara, K. Development of a Deep Learning-Based Algorithm to Detect the Distal End of a Surgical Instrument. *Appl. Sci.* **2020**, *10*, 4245. [CrossRef]
13. Koskinen, J.; Torkamani-Azar, M.; Hussein, A.; Huotarinen, A.; Bednarik, R. Automated Tool Detection with Deep Learning for Monitoring Kinematics and Eye-Hand Coordination in Microsurgery. *Comput. Biol. Med.* **2022**, *141*, 105121. [CrossRef] [PubMed]
14. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
15. Redmon, J.; Farhadi, A. Yolo9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
16. Redmon, J.; Farhadi, A. Yolov3: An Incremental Improvement. Available online: https://arxiv.org/pdf/1804.02767.pdf (accessed on 17 February 2023).
17. Glenn, J.; Alex, S.; Jirka, B. Ultralytics/Yolov5. Available online: https://github.com/ultralytics/Yolov5 (accessed on 17 February 2023).
18. Zhang, Y.; Guo, Z.; Wu, J.; Tian, Y.; Tang, H.; Guo, X. Real-Time Vehicle Detection Based on Improved YOLO v5. *Sustainability* **2022**, *14*, 12274. [CrossRef]
19. Kim, J.-H.; Kim, N.; Park, Y.W.; Won, C.S. Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [CrossRef]
20. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
21. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia (MM '16), Association for Computing Machinery, New York, NY, USA, 15–19 October 2016; pp. 516–520.
22. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
23. Tong, Z.; Chen, Y. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. Available online: https://arxiv.org/pdf/2301.10051.pdf (accessed on 24 January 2023).
24. Gennari, M.; Fawcett, R. DSConv: Efficient Convolution Operator. Available online: https://arxiv.org/pdf/1901.01928.pdf (accessed on 7 January 2019).