symmetry

MDPI

Article

Semantic SLAM Based on Deep Learning in Endocavity Environment

Haibin Wu [1], Jianbo Zhao [1], Kaiyang Xu [1], Yan Zhang [1], Ruotong Xu [1], Aili Wang [1,*] and Yuji Iwahori [2]

[1] Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China; woo@hrbust.edu.cn (H.W.); 2020600002@stu.hrbust.edu.cn (J.Z.); 1920610061@stu.hrbust.edu.cn (K.X.); 2020610056@stu.hrbust.edu.cn (Y.Z.); 1920610059@stu.hrbust.edu.cn (R.X.)

[2] Department of Computer Science, Chubu University, Aichi 487-8501, Japan; iwahori@isc.chubu.ac.jp

* Correspondence: aili925@hrbust.edu.cn

Abstract: Traditional endoscopic treatment methods restrict the surgeon's field of view. New approaches to laparoscopic visualization have emerged due to the advent of robot-assisted surgical techniques. Lumen simultaneous localization and mapping (SLAM) technology can use the image sequence taken by the endoscope to estimate the pose of the endoscope and reconstruct the lumen scene in minimally invasive surgery. This technology gives the surgeon better visual perception and is the basis for the development of surgical navigation systems as well as medical augmented reality. However, the movement of surgical instruments in the internal cavity can interfere with the SLAM algorithm, and the feature points extracted from the surgical instruments may cause errors. Therefore, we propose a modified endocavity SLAM method combined with deep learning semantic segmentation that introduces a convolution neural network based on U-Net architecture with a symmetric encoder–decoder structure in the visual odometry with the goals of solving the binary segmentation problem between surgical instruments and the lumen background and distinguishing dynamic feature points. Its segmentation performance is improved by using pretrained encoders on the network model to obtain more accurate pixel-level instrument segmentation. In this setting, the semantic segmentation is used to reject the feature points on the surgical instruments and reduce the impact caused by dynamic surgical instruments. This can provide more stable and accurate mapping results compared to ordinary SLAM systems.

Keywords: augmented reality; simultaneous localization and mapping (SLAM); deep learning; semantic segmentation

Citation: Wu, H.; Zhao, J.; Xu, K.; Zhang, Y.; Xu, R.; Wang, A.; Iwahori, Y. Semantic SLAM Based on Deep Learning in Endocavity Environment. *Symmetry* **2022**, *14*, 614. https://doi.org/10.3390/sym14030614

Academic Editor: Theodore E. Simos

Received: 6 January 2022
Accepted: 8 February 2022
Published: 19 March 2022

1. Introduction

Most hospitals are now equipped with two-dimensional endoscopes to assist doctors in minimally invasive surgery of the abdominal cavity, thoracic cavity, and throat. The advantage is that doctors no longer need to cut the abdominal cavity and can operate only through a tiny incision in the abdomen. Compared to traditional surgery or earlier minimally invasive surgery, modern minimally invasive surgery has the advantages of smaller incisions, less bleeding, and faster recovery of patients after surgery, all of which reduce patient trauma and pain. Therefore, it is increasingly popular and widely used in surgery [1]. However, surgeons are prone to disorientation and hand–eye dissonance while finding targets and performing complex procedures through the 2D visual display of the endoscopic video stream, making it difficult to empirically match the laparoscopic field of view with the preoperative images to determine the site of the lesion. The main reason for this is the relatively incomplete or poor visual feedback, which does not allow direct observation of the overall environment of the internal cavity. Reconstructing the three-dimensional (3D) image of the lesion site from the acquired images enables the surgeon to make a more accurate diagnosis.

Minimally invasive surgery is gradually merging with computer vision techniques, and the use of computers for image processing to extend the surgical field of view has broken the limitations of traditional surgery [2–4]. Three-dimensional models allow a more intuitive view of the surgical scene and simplify the localization process. There are some solutions for the three-dimensional imaging methods based on computer vision in the lumen environment. For example, time-of-flight and structured light-based solutions have also been used for dense scene reconstruction, but they are less commonly used in the endocavity environment due to hardware constraints such as size and cost [5–7]. Structure-from-Motion (SFM) has been widely used in studies relating to camera motion tracking and 3D reconstruction [8,9], but it is ineffective for low-resolution, weakly textured images such as internal cavities and requires offline processing; hence, it has a limited application in internal cavities. Shape-from-Shading (SFS) performs 3D reconstruction of internal cavity organs based on light and dark changes in image grayscale information [10], but there are multiple mappings between 2D images and surface shapes, resulting in relatively higher errors in the results obtained by SFS.

In recent years, simultaneous localization and mapping, such as Simultaneous Localization and Mapping from Visual sensor (VSLAM), used only the camera as an external sensor for robotics self-localization in unknown environments and reconstructing maps of the surroundings [11]. VSLAM has also been gradually applied to the medical field to provide new solutions for 3D reconstruction of internal cavities by moving the endoscope and using the features of each image frame for incremental composition of soft tissues and organ structures. Such methods have been used for endoscopic image reconstruction in [12–15].

Nonetheless, in VSLAM, if dynamic objects are present in the real scene, the moving objects cause degradation in the performance and robustness of the SLAM system and may lead to errors in image feature matching, camera pose estimation, map construction, and loop closure detection, thereby making the algorithm fail. Therefore, recognition and removal of dynamic objects in the scene needs to be considered. The first system to successfully fuse SLAM with moving object detection and tracking in a normal scene was proposed by Wang [16], and subsequently several researchers have investigated SLAM in dynamic environments. For example, Kundu [17] detected dynamic features by using epipolar geometric constraints and flow vector bound (FVB) constraints, but the effect of rejecting dynamic feature points is affected by the accuracy of the visual odometry, and it is not able to properly handle a moving object that stops in the middle. Wang [18] implemented moving object segmentation based on optical flow computation followed by sparse point trajectory clustering and densification, but the optical flow method works on the basis of constant light brightness, so it is easily affected by light changes. Moreover, some researchers have also tried to combine deep learning to remove potential dynamic objects by methods such as semantic segmentation, e.g., a complete semantic SLAM system in dynamic environments (DS-SLAM) [19] used a deep convolutional encoder-decoder architecture for image segmentation (SegNet) combined with a movement consistency detection method in order to filter out moving people in the scene, but the improvements in low dynamic sequences were small and humans were the only dynamic objects targeted. Tracking, mapping and inpainting by SLAM in dynamic scenes (DynaSLAM) [20] removes dynamic objects, such as people and cars, by performing pixel-by-pixel segmentation of a-priori dynamic objects in images using the instance segmentation network called region-based masked convolutional network (Mask R-CNN). This method chooses to remove all potentially moving objects, which may leave too few static feature points available to affect the camera's pose estimation. The method is also computationally intensive and time-consuming.

In minimally invasive surgery, the only typically dynamic objects that have an impact are surgical instruments, such as surgical forceps and scissors, without fear of sudden movement of other potential objects. Nevertheless, due to the change of light in the lumen and the complexity and dynamics of the background tissue, the segmentation

of surgical instruments is more difficult. Related applications of deep-learning-based methods to robotic instrument segmentation proved their excellent performance in internal cavity binary and multiclass segmentation [21,22]. Therefore, we considered combining deep learning for surgical instrument segmentation in the SLAM process to prevent the mismatching of extracted features and other situations.

In medical image segmentation, U-Net [23] shows very excellent performance, as it is able to build more feature channels in the upsampling phase while associating feature maps from the downsampling phase due to the presence of skip connections. In addition, the model training requires fewer datasets and is able to converge on a small amount of data. U-Net has now become a cornerstone in the field of medical image segmentation [24] and is the most popular and effective technique for dealing with medical image segmentation. For example, Francia et al. added residual blocks to the U-Net network to achieve accurate segmentation of retinal blood vessels [25]. Ding et al. proposed a U-Net-based deep attention network with a color normalization operation to implement end-to-end segmentation of the glottal region [26]. Siddique et al. reviewed the application of U-Net in the field of medical image segmentation and in other medical image analyses and pointed out that the U-Net based architecture is quite innovative and valuable to medical image diagnosis and is one of the most important deep learning techniques [27].

In this paper, we note that although visual SLAM is widely used in medical scenarios, it is fragile and difficult to use stably under dynamic environments. Especially in the minimally invasive surgery setting, the impact of moving surgical instruments is a problem that needs to be considered for medical robots facing complex environments in internal cavities. Deep learning allows intelligent segmentation of images in order to understand what is in them and makes analysis of each part easier. Detecting and processing dynamic objects is necessary for SLAM to estimate a stable map, which is helpful for its application. From this perspective, we hypothesize that the use of application-specific CNN networks in SLAM systems to reject dynamic objects can reduce the false association information when the SLAM system works. For neural networks, better performance can be obtained by using pretrained encoders, which also help to avoid overfitting.

Accordingly, we propose the use of a neural network based on the U-Net architecture to distinguish dynamic features on surgical instruments from static features in the background by segmenting the surgical instruments, rejecting the dynamic features as outliers, and using only the static features to track the endoscope position as well as to complete the subsequent SLAM process. Thus, semantic SLAM based on deep learning in the endocavity environment is proposed in this paper. By incorporating semantic segmentation, a dynamic SLAM system is constructed that can operate in complex internal cavity environments. The moving surgical instruments are masked out using semantic information. This solves the problem of feature extraction, localization, and 3D reconstruction in the case of moving surgical instruments in the internal cavity. Avoiding errors caused by incorrect matching and system crashes during internal cavity SLAM enables more robust SLAM performance in an inner cavity environment.

The rest of this paper is organized as follows: Section 2 introduces the proposed SLAM and deep learning work. Section 3 gives the experimental results and analysis. Finally, the conclusion is made in Section 4.

2. Materials and Methods

The overall framework of the proposed dynamic SLAM system for the internal cavity is shown in Figure 1. The SLAM system based on the Oriented FAST and Rotated Brief (ORB-SLAM) [28] consists of four main modules: instrument segmentation, tracking, local mapping, and loop closing. We introduced semantic segmentation based on convolutional neural networks into the SLAM process to construct binary masks from RGB images, ensuring dynamic feature points on the surgical instruments were eliminated in the tracking module to avoid incorrect data association information. Then, only the feature points extracted from the regions other than the surgical instruments were used for localization

and map construction of the internal cavity scene. Thus, the SLAM system could estimate the endoscope motion more consistently, as well as obtain accurate soft tissue reconstruction of the internal cavity. It is possible to give the surgeon more intuitive visual feedback in the endoscopic SLAM system, which can assist the surgeon in making judgements. In order to facilitate the possible manipulation of surgical instruments while visualizing the endoscopic scene, the influence of dynamic objects, i.e., surgical instruments, on the system was significantly reduced by combining segmentation networks.

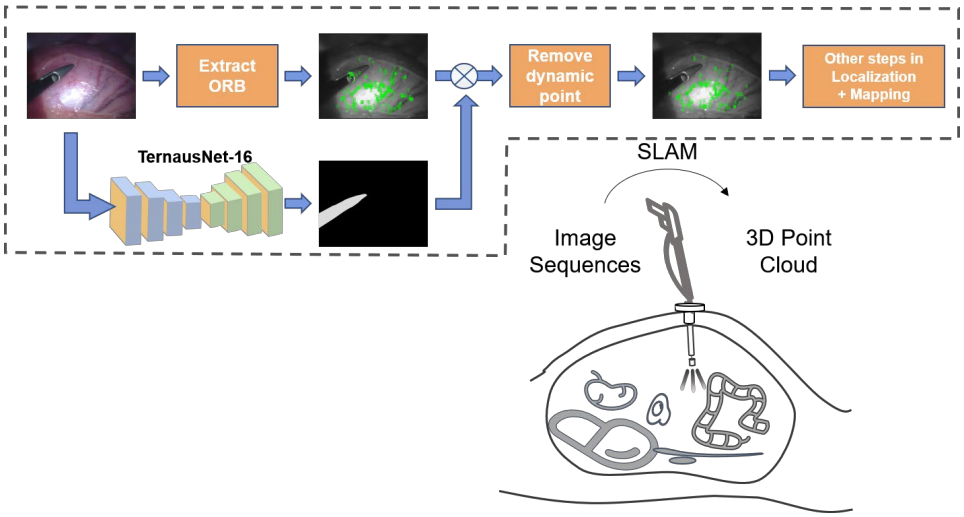


Figure 1. System framework of SLAM combined with semantic segmentation.

The endoscope is first used to observe the environment of the inflated internal cavity, and the acquired sequence of internal cavity images is transmitted to the SLAM tracking thread and the semantic segmentation network based on the U-Net architecture. Since it is in a minimally invasive surgery scene, moving medical instruments are considered in the image frames, and the surgical instruments are segmented by pixel through the semantic segmentation network to obtain a binary mask. In the tracking thread, for each new frame, ORB feature points with stable geometric features are extracted. The mask obtained from the segmentation is also used to judge whether the current feature point is a dynamic feature point. If so, it is identified as an outlier and eliminated. Otherwise, it can be classified as a static candidate feature point. The static feature points are then used for the subsequent tracking and mapping steps of SLAM, including estimating the camera pose using the matching correspondence of adjacent frames, obtaining the depth estimation using the triangulation method, and jointly optimizing the map and camera pose using local and global bundle adjustments.

2.1. Surgical Instrument Segmentation

In addition to the moving surgical instruments in the endocavity scene, the surrounding environment is not completely rigid and the organs or soft tissues are also subject to certain deformations, making it very difficult to distinguish moving objects from the scene using only the geometric approach in SLAM. To separate the surgical instruments from the soft tissue background, we used a neural network based on the U-Net architecture to segment the surgical instruments using semantic information.

U-Net is a fully convolutional network with a symmetric encoder–decoder structure. This encoder–decoder structure of U-Net containing skip connections is able to fuse features from different layers to obtain accurate pixel-level localization with excellent segmentation results and was shown to perform well for segmentation problems with limited data [29], making it well-suited for segmentation of surgical instruments in medical scenarios. To improve its binary segmentation performance, we used a pretrained VGG16 encoder on the U-Net infrastructure, which is called TernausNet-16 [30], and then integrated it into SLAM

as the semantic segmentation network of the system. Figure 2 illustrates the TernausNet-16 network model used in the proposed segmentation algorithm, which is a classical full convolution network. Each rectangular box represents a transformed multichannel feature map. The number of channels is below the rectangular box. The height of the box corresponds to the resolution of the different feature maps. The blue arrows indicate skip connections, where information is transferred from the encoder to the decoder. The lumen image is used as the input. The left side is the encoder for downsampling, also known as the contracting path, and the right side is the decoder for upsampling, also known as the expansive path. The two are associated through skip connection.

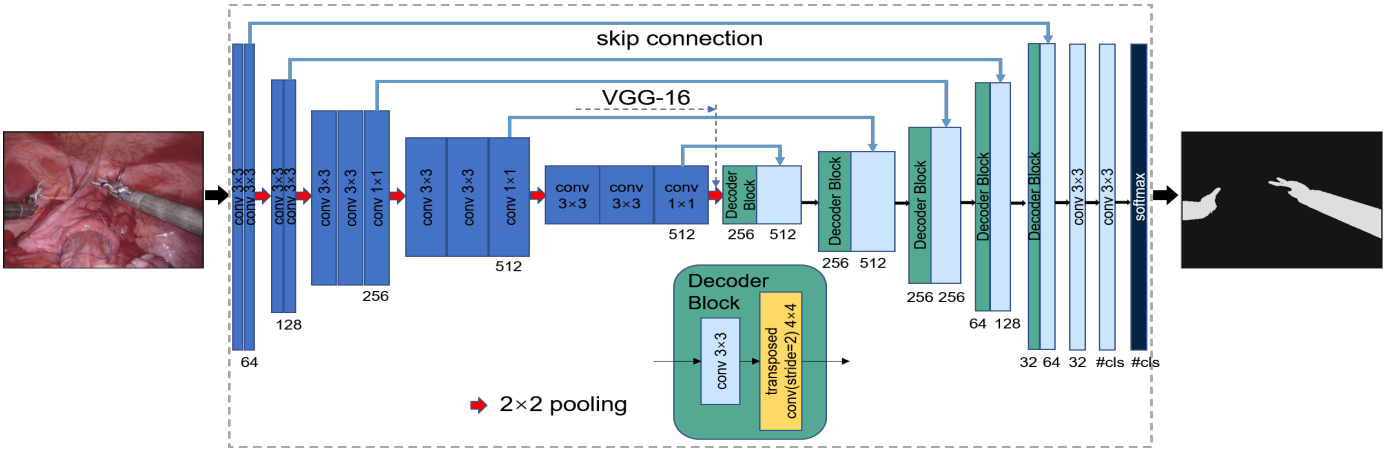


Figure 2. The structure of the TernausNet-16 Network Model.

The specific operation of the encoder for this network is given in Table 1; a simple pre-trained VGG16 network is used as the encoder. VGG16 consists of 16 forward-propagating network layers, which contain 13 convolutional layers. The convolutional kernel size is 3×3, and each convolutional layer is immediately followed by an ReLU activation function. The convolution layer is also followed by five 2×2 max pooling layers, which perform dimensionality reduction on the feature map. The first convolutional layer has 64 channels, and each subsequent pooling operation doubles the number of channels up to 512.

Table 1. Encoder Configuration.

Operator	Size	Filter	Layers
Convolution_1	3×3	64	2
Maxpool_1	$2 \times 2/2$	\	1
Convolution_2	3×3	128	2
Maxpool_2	$2 \times 2/2$	\	1
Convolution_3	3×3	256	2
Convolution_4	1×1	256	1
Maxpool_3	$2 \times 2/2$	\	1
Convolution_5	3×3	512	2
Convolution_6	1×1	512	1
Maxpool_4	$2 \times 2/2$	\	1
Convolution_7	3×3	512	2
Convolution_8	1×1	512	1
Maxpool_5	$2 \times 2/2$	\	1

The specific architecture of the decoder is given in Table 2. The decoder section is a symmetrical structure to the encoder, and replaces the fully connected layer with a convolutional layer. The feature map size is enlarged by using transposed convolution. The output of the transposed convolution is used as the output feature map of the corresponding

part of the decoder, which is then processed by direct convolution operations to keep the number of channels the same as the symmetric encoder term. At the end of the network, the feature maps of the background and target foreground are obtained, and then a probability map of the categories is obtained by the soft-max function.

Table 2. Decoder Configuration.

Operator	Size	Filter	Layers
Convolution_9	3×3	512	1
Transposed Convolution_1	$4 \times 4/2$	256	1
Convolution_10	3×3	512	2
Transposed Convolution_2	$4 \times 4/2$	256	1
Convolution_11	3×3	512	2
Transposed Convolution_3	$4 \times 4/2$	256	1
Convolution_12	3×3	256	2
Transposed Convolution_4	$4 \times 4/2$	64	1
Convolution_13	3×3	128	2
Transposed Convolution_5	$4 \times 4/2$	32	1
Convolution_14	3×3	64	2
Soft-max	\	\	1

Since the network combines low-resolution information in downsampling and high-resolution information in upsampling and shallow and deep features of the image, it is able to achieve excellent pixel-by-pixel localization and segmentation. As an output of the model, the surgical instruments are distinguished from the pixel values in the background area, and a binary mask can be obtained by binarizing the pixel probabilities finally.

2.2. Internal Cavity Vision SLAM

ORB-SLAM2 is an advanced visual SLAM system based on feature tracking that has reliable and excellent performance in most scenarios. There have been related works applying ORB-SLAM2 to 3D reconstruction of internal cavities [31,32] which proved that it can cope with the complex environment of internal cavities. Therefore, this paper implements a global feature-based endoluminal SLAM scheme based on ORB-SLAM2 and improves the robustness and accuracy of the endoluminal SLAM system by adding a decision module that uses semantic information to distinguish surgical instruments, segmenting dynamic features and removing them as outliers. A brief framework of the system proposed in this paper is plotted in Figure 3.

The system performs ORB feature point extraction for each image frame collected by the endoscope in the tracking thread, compares the descriptors of each feature point, thus obtaining the corresponding point pairs, and then estimates the endoscope motion based on the correspondence. Therefore, the correctness of feature points and their matching relationships are important for the tracking and mapping results of SLAM. Random sample consistency (RANSAC) is an algorithm that estimates the parameters of a model in an iterative manner in and obtains valid data from data containing outliers. It is usually used to eliminate outliers from a large number of matched point pairs and select the more reliable pairs. However, the dynamic feature points on the surgical instruments can also produce incorrect matching relationships, and the probability of the extracted

feature points becoming internal points will gradually increase when the moving surgical instruments appear in the picture for a longer period of time. Therefore, we used the mask obtained by semantic segmentation to accurately reject the dynamic feature points on the surgical instruments.

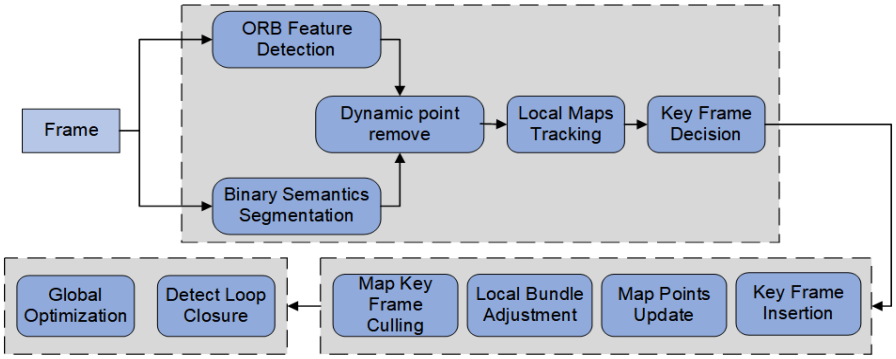


Figure 3. Brief framework diagram of internal cavity vision SLAM proposed in this paper.

Based on the initially extracted ORB feature points and the binary mask results obtained through the TernausNet-16 segmentation network, the preselected feature points were removed if they were within the mask range, thus excluding the erroneous feature points detected on the moving surgical instruments. Suppose the set of feature points extracted from the input Kth frame image is given by the following equation

$$A_l^K = \left\{ p_1^K, p_2^K, p_3^K, \ldots, p_i^K \right\}, \tag{1}$$

where p_i^K is the ith feature point in the K^{th} frame. The set of pixels in the region where the surgical instruments were located in the binary mask of the K^{th} frame image is defined as follows:

$$S_N^K = \left\{ s_1^K, s_2^K, s_3^K, \ldots, s_n^K \right\}, \tag{2}$$

where s represents the pixel point in the area where the mask is located in the frame. If there was a point in the sequence of feature points that satisfied $p_i^K \in S_N^K$, it would be identified as a dynamic feature point and removed from the sequence of feature points.

By using the binary mask generated by semantic segmentation, preselected feature points were filtered and dynamic feature points located on the surgical instruments were successfully removed, thus evading the detrimental effect of incorrect correspondence on SLAM. At the same time, feature points in other background regions were used as static feature points, and then outliers were further removed by the RANSAC algorithm. Finally, the endoscope's pose was estimated based on the correct correspondence.

Once the initial pose estimation was completed, the subsequent estimation could be continued by the Perspective-n-Point (PnP) algorithm or Iterative Closest Point (ICP) algorithm. The match between the current frame and the local map was obtained by tracking the endoscopic pose and the local map. Pose optimization was performed using minimization of reprojection errors, and then the keyframe generation is determined by the pose and motion of adjacent frames. In the local mapping thread, the local map is constructed by filtering the newly generated map points, triangulating the map points with a high degree of coviewing, performing Bundle Adjustment (BA) optimization, and removing redundant keyframes. Lastly, the map was updated by performing global BA optimization on the global poses and map points.

3. Experimental Results and Discussions

This paper constructs a modified endoscopic SLAM algorithm combined with semantic segmentation (SS-SLAM), using the Hamlyn Center's endoscopic video dataset (London,

UK) [33] to validate the overall construction improvement. This dataset includes endoscopic scene images of various organs and soft tissues for tasks such as polyp detection, image segmentation, and localization. The sequences with instrumental invasion were selected as experimental data. To verify that the proposed method in this paper could effectively reject dynamic feature points on surgical instruments when performing monocular SLAM in an internal cavity scene, experiments were conducted using publicly available medical image datasets and compared to the original ORB-SLAM2 algorithm. The experiments were all conducted on a computer equipped with an Intel Xeon D-1581 CPU, NVIDIA GTX 1070Ti GPU, and 32 G RAM.

The segmentation network proposed above was first trained in order to segment a-priori moving objects, i.e., surgical instruments, under endoscopic images using frame sequences acquired from the da Vinci Surgical System provided by the MICCAI [34] in Quebec, Canada. Each image was in RGB format and had a 1920 $\times$ 1080-pixel resolution. The training dataset had 8 $\times$ 255 frame sequences. True value labels were provided for each image frame in the dataset, and the labels of the various parts of the surgical instruments were manually labelled in each frame for training purposes. The frames in each video were correlated, so we performed 4-fold cross-validation and split the data based on this dependency, dividing the training set into four quarters. Three quarters were used for training and one for validation. We repeated this four times until all quarters had a chance to be the validation set at least once. The network was trained using the Adam optimizer for 40 epochs, with the initial learning rate set to 0.00001. The original RGB images were passed into the system and a pixel-by-pixel prediction mask of the images was obtained after segmentation of the surgical instruments present in the images by the TernausNet-16 network to achieve a binary semantic segmentation of the surgical instruments and the background.

The Jaccard index, also referred to as the Intersection Over Union (IoU), was chosen for the evaluation metric, which was used to measure the similarity between two sets. For two finite sets A and B, it is defined as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{3}$$

For images, the above equation can also be rewritten in the following form:

$$J = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i \hat{y}_i}{y_i + \hat{y}_i - y_i \hat{y}_i}\right), \tag{4}$$

where y_i is the binary label of pixel i and $\hat{y}_i$ is its predicted probability. For the image binary segmentation problem, which can also be viewed as a pixel classification problem, a combination of the binary cross-entropy loss function H is used in this paper as follows:

$$L = H - \log J. \tag{5}$$

U-Net achieves excellent performance in different biomedical segmentation applications and is a very classical network. The validation learning curve of the TernausNet-16 network versus the original U-Net network is represented in Figure 4. The TernausNet-16 network model converges to a stable value much faster than the original U-Net. The final steady-state result of TernausNet-16 is also higher than that of the U-Net, which means that the improved segmentation framework performs better segmentation of surgical instruments with higher accuracy compared to the original U-Net network.

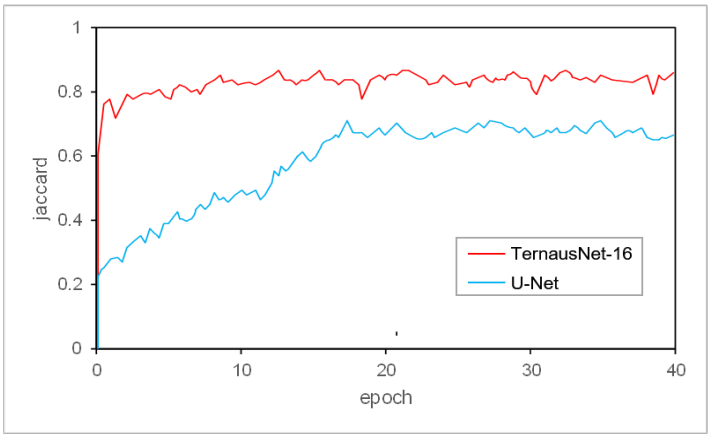


Figure 4. Validation learning curves of different network models.

Another commonly used metric is the Dice coefficient. Given a vector of ground truth labels T_1 and a vector of predicted labels P_1, Dice coefficient can be defined as:

$$D(T_1, P_1) = \frac{2|T_1 \cap P_1|}{|T_1| + |P_1|}. \tag{6}$$

The Dice coefficient is usually used to calculate the similarity of two sets, with values ranging from zero to one. The performance characterization of segmentation algorithms for different segmentation targets is an important issue. It is therefore more reasonable to evaluate the accuracy of the proposed method through multiple performance metrics.

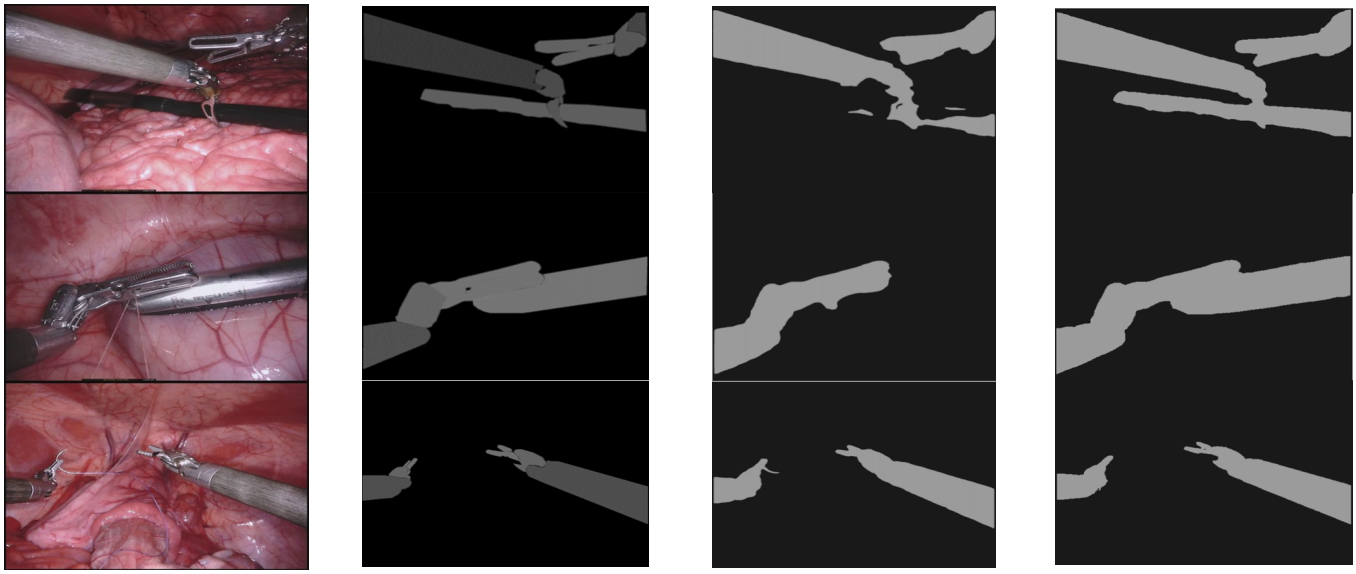
Specifically relating to the training accuracy of binary segmentation, U-Net obtained an IoU of 0.721 while TernausNet-16 obtained an IoU of 0.842, which showed an improvement in training accuracy. Furthermore, a quantitative comparison was carried out on the test set and the results are shown in Table 3. Predictions were made for each image and the final results were averaged. Our model achieved better results; its IoU was 0.813 in comparison with an IoU of 0.698 for the U-Net and its Dice coefficient was 0.894 in comparison with a Dice coefficient of 0.805 for the U-Net. We conducted statistical tests to compare the performance of the segmentation in terms of IoU and Dice metrics. Using the Wilcoxon Signed Rank Test, TernausNet-6 was found to display statistically significant improvement ($p < 0.05$) in IoU and Dice over the U-Net.

Table 3. The quantitative results of segmentation by different network models.

Network	IoU	Dice Coefficient
U-Net	0.708	0.805
TernausNet-16	0.826	0.894

Figure 5 shows the results of the segmentation of moving surgical instruments. The segmentation results clearly show that our model based on TernausNet-16 was able to segment the surgical instruments more completely and the segmented images were closer to the ground truth. However, the segmentation results of U-Net have some omissions and mislabeling. This indicates that the dynamic surgical instruments can be detected more accurately by our segmentation network and the corresponding binary masks can be generated. The input monocular RGB images are preprocessed where the surgical instruments are segmented in order to facilitate the subsequent rejection of dynamic feature points extracted in SLAM. The mask obtained from the above results can cover the area where the surgical instruments are located, so it can be used for subsequent processing to properly remove the feature points extracted from this part of the area. Overall, the performance of the U-Net model was improved by adjusting the encoder part. In the binary segmentation task, it could converge to the optimal stable value more quickly and reduce

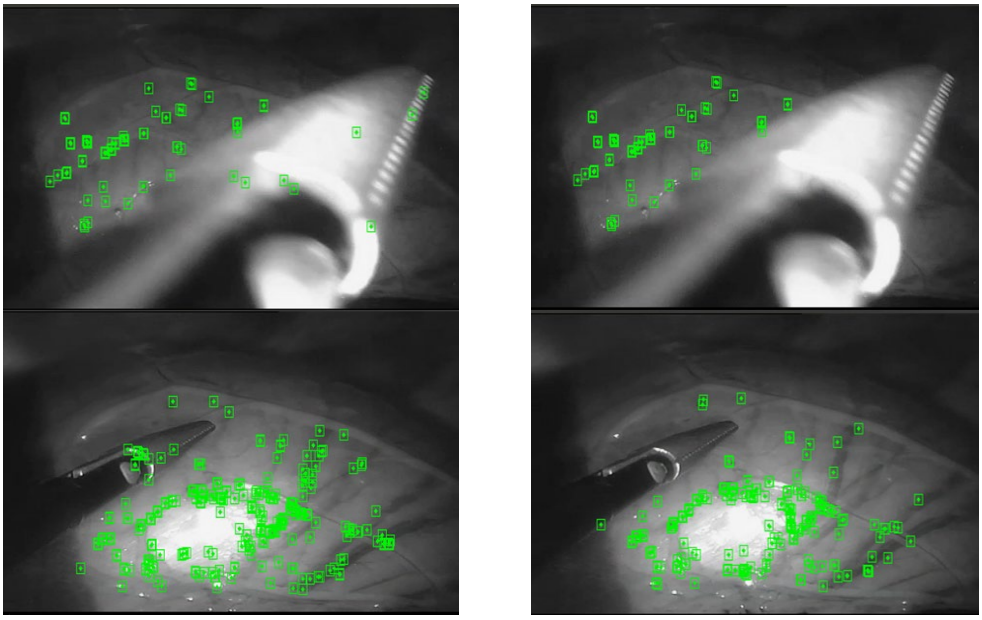
the training time of the model. The final accuracy has also been improved, allowing for a more detailed and complete segmentation profile of the surgical instrument.



(**a**) (**b**) (**c**) (**d**)

Figure 5. Comparison of semantic segmentation results for surgical instruments. (**a**) The original image; (**b**) The ground truth; (**c**) U-Net-based segmentation; (**d**) TernausNet-16-based segmentation.

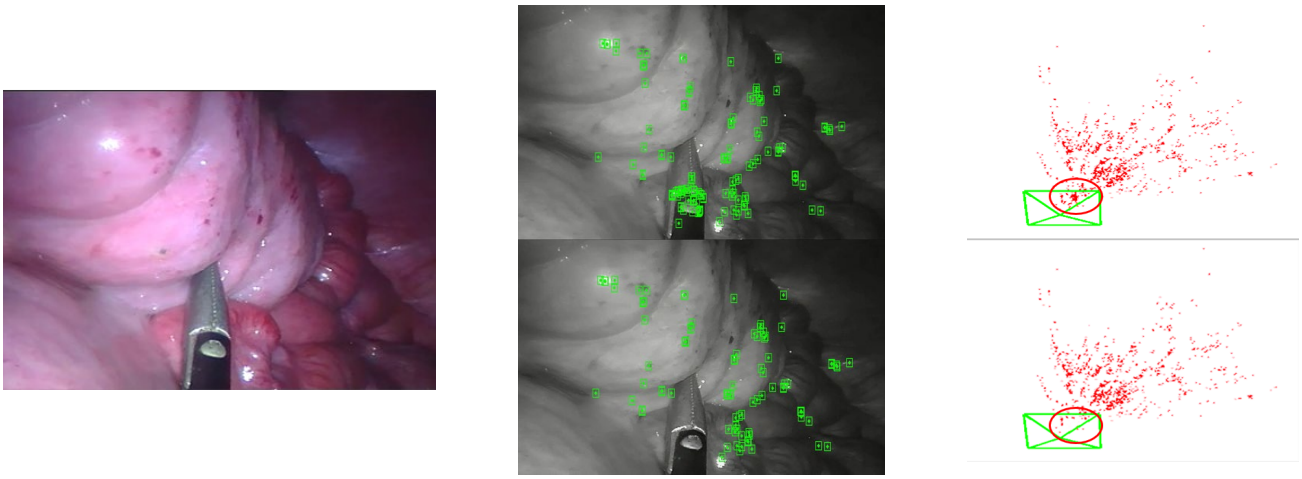
For each input frame containing surgical instruments, a mask is obtained by binary semantic segmentation calculation. The features obtained by the ORB feature extraction algorithm in the SLAM system are removed from the feature sequence when they are located in the mask region, while the features in other regions continue to be used for subsequent tracking and mapping. By using the mask to limit the feature detection area and thus prevent the feature points from concentrating on the surgical instruments, false extraction and matching can be avoided. As shown in Figure 6, the feature point detection on the surgical instruments is successfully excluded by using the mask.



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

Figure 6. Comparison of feature extraction results by different methods. (**a**) SLAM; (**b**) SS-SLAM.

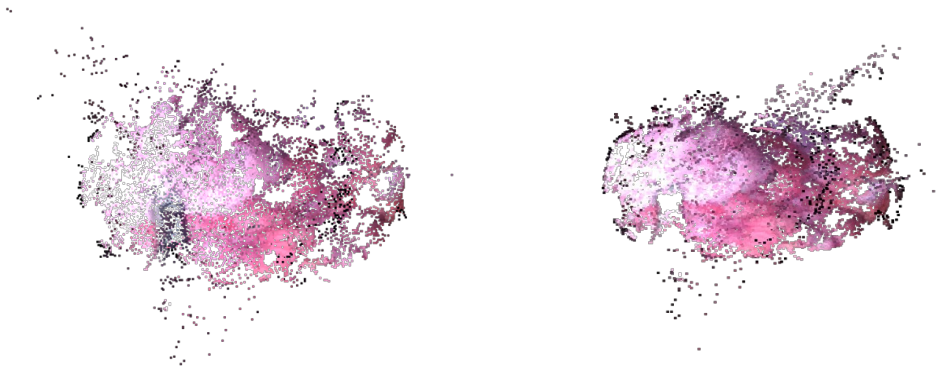
Compared with the original ORB-SLAM2 algorithm, the feature points on the surface of dynamic surgical instruments are eliminated using semantic segmentation, and a higher quality of map construction can be obtained. In Figure 7, the feature points on the surgical instruments were successfully excluded in the tracking and mapping.



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

Figure 7. Results of endoluminal SLAM reconstruction. (**a**) frame from datasets video; (**b**) feature points extraction results obtained by SLAM (top) and SS-SLAM (bottom); (**c**) 3D points obtained by SLAM (top) and SS-SLAM (bottom).

The more detailed visualization results obtained after extracting more feature points and performing the densification operation are shown in Figure 8, where our proposed method yields more accurate results in reconstructing the map points and successfully rejects the surgical instruments in the region where the surgical instruments move for a long time. It can be seen in the figure that there are more missing parts of the reconstruction results in (a), probably due to dynamic interference, but our method results in a better reconstruction of the inner cavity background by removing dynamic feature points. Consequently, our method has significantly improved the reconstruction of the internal cavity background, which is beneficial to improving the accuracy of 3D reconstruction by monocular visual SLAM.



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

Figure 8. Results of reconstructed dense point cloud. (**a**) SLAM; (**b**) SS-SLAM.

4. Conclusions

In this paper, we build an endoluminal vision SLAM framework incorporating semantic segmentation for minimally invasive surgery scenarios. The semantic segmentation network was based on the TernausNet-16 network architecture and improved in order to

effectively segment the surgical instruments in the internal cavity image. Then a dynamic feature point judgment module was added to the SLAM to remove the feature points in the surgical instrument mask region, thus eliminating the dynamic feature points detected on the surgical instruments and using only reliable feature points to provide a good basis for the subsequent modules. This enabled the SLAM system to be more robust in processing internal cavity image sequences and to obtain more accurate mapping results. In the endoscopic video dataset experiments, our proposed method achieved better results in both surgical instrument segmentation and mapping. In the future, extensions of this work may include building effective models of endoluminal soft tissue deformation to cope with more complex endoluminal scenarios and considering densification operations on less endoluminal data to eventually achieve a realistic dense reconstruction of the endocavity.

Author Contributions: Conceptualization, H.W. and A.W.; Methodology, J.Z., K.X., Y.Z. and R.X.; Software, J.Z., K.X., Y.Z. and R.X.; Supervision, Y.I.; Validation, J.Z., K.X., Y.Z. and R.X.; Writing—review & editing, H.W. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China under Grant NSFC-61671190.

Data Availability Statement: The data presented in this study are available on http://hamlyn.doc.ic.ac.uk/vision/ (accessed on 1 November 2021).

Acknowledgments: We thank Kaiyuan Jiang for his valuable comments and discussion.

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

References

1. Jang, J.Y.; Han, H.S.; Yoon, Y.S.; Cho, J.Y.; Choi, Y.R. Retrospective comparison of outcomes of laparoscopic and open surgery for t2 gallbladder cancer—Thirteen-year experience. *Surg. Oncol.* **2019**, *29*, 142–147. [CrossRef] [PubMed]
2. Totz, J.; Fujii, K.; Mountney, P.; Yang, G.Z. Enhanced visualisation for minimally invasive surgery. *Int. J. Comput. Assist. Radiol. Surg.* **2012**, *7*, 423–432. [CrossRef] [PubMed]
3. Vemuri, A.S.; Liu, K.C.; Ho, Y.; Wu, H.S.; Ku, M.C. Endoscopic Video Mosaicing: Application to Surgery and Diagnostics. In Proceedings of the Living Imaging Workshop, Strasbourg, France, 1 December 2012.
4. Afifi, A.; Takada, C.; Yoshimura, Y.; Nakaguchi, T. Real-time expanded field-of-view for minimally invasive surgery using multi-camera visual simultaneous localization and mapping. *Sensors* **2021**, *21*, 2106. [CrossRef] [PubMed]
5. Brandt, O.; Munwes, Y. Commissioning and First Image Reconstruction with a New Time-of-Flight PET Prototype. In Proceedings of the 2018 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Sydney, Australia, 1 November 2018.
6. Furukawa, R.; Oka, S.; Kotachi, T.; Okamoto, Y.; Tanaka, S.; Sagawa, R.; Kawasaki, H. Fully Auto-calibrated Active-stereo-based 3D Endoscopic System using Correspondence Estimation with Graph Convolutional Network. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 4357–4360.
7. Zorraquino, C.; Bugalho, R.; Rolo, M.D.D.R.; Silva, J.C.; Vecklans, V.; Silva, R.; Ortigao, C.; Neves, J.A.; Tavernier, S.; Guerra, P.; et al. Asymmetric data acquisition system for an endoscopic pet-us detector. *IEEE Trans. Nucl. Sci.* **2016**, *63*, 213–221. [CrossRef]
8. Sun, D.; Liu, J.; Linte, C.A.; Duan, H.; Robb, R.A. Surface Reconstruction from Tracked Endoscopic Video Using the Structure from Motion Approach. In *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*; Springer: Berlin/Heidelberg, Germany, 2013.
9. Wei, L.; Dong, M.-L.; Lu, N.-G.; Lou, X.-P. Hand-eye calibration method without a calibration reference based on second-order cone programming. *Opt. Precis. Eng.* **2018**, *26*, 2536–2545. [CrossRef]
10. Collins, T.; Bartoli, A. Towards Live Monocular 3D Laparoscopy Using Shading and Specularity Information. In Proceedings of the International Conference on Information Processing in Computer-Assisted Interventions, Pisa, Italy, 27 June 2012.
11. Qi, H.; Wen-long, C.; Ding-fan, L.; Min, J. Survey on Monocular Visual Inertial SLAM Algorithms. *Softw. Guide* **2020**, *19*, 6. (In Chinese)
12. Mountney, P.; Yang, G.Z. Dynamic view expansion for minimally invasive surgery using simultaneous localization and mapping. In Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Minneapolis, MN, USA, 3–6 September 2009; pp. 1184–1187.
13. Lin, B.; Johnson, A.; Qian, X.; Sanchez, J.; Sun, Y. Simultaneous tracking, 3D reconstruction and deforming point detection for stereoscope guided surgery. In *Augmented Reality Environments for Medical Imaging and Computer-Assisted Interventions*; Springer: Berlin/Heidelberg, Germany, 2013.

14. Xie, C.; Yao, T.; Wang, J.; Liu, Q. Endoscope localization and gastrointestinal feature map construction based on monocular SLAM technology. *J. Infect. Public Health* **2020**, *13*, 1314–1321. [CrossRef] [PubMed]
15. Peng, X. Research on Endoscopic Visual SLAM for Minimally Invasive Surgery. Master's Thesis, University of Electronic Science and Technology of China, Chengdu, China, 20 June 2017.
16. Wang, C.; Thorpe, C. Simultaneous localization and mapping with detection and tracking of moving objects. In Proceedings of the 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292), Washington, DC, USA, 11–15 May 2002; Volume 3, pp. 2918–2924.
17. Kundu, A.; Krishna, K.M.; Sivaswamy, J. Moving object detection by multi-view geometric techniques from a single camera mounted robot. In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, St. Louis, MO, USA, 10–15 October 2009.
18. Wang, Y.; Huang, S. Towards dense moving object segmentation based robust dense RGB-D SLAM in dynamic scenarios. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014.
19. Yu, C.; Liu, Z.; Liu, X.-J.; Xie, F.; Yang, Y.; Wei, Q.; Fei, Q. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018.
20. Bescos, B.; Fácil, J.M.; Civera, J.; Neira, J. DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes. *IEEE Robot. Autom. Lett.* **2018**, *3*, 4076–4083. [CrossRef]
21. Laina, I.; Rieke, N.; Rupprecht, C.; Vizcaíno, J.P.; Eslami, A.; Tombari, F.; Navab, N. Concurrent Segmentation and Localization for Tracking of Surgical Instruments. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2017.
22. Pakhomov, D.; Premachandran, V.; Allan, M.; Azizian, M.; Navab, N. Deep residual learning for instrument segmentation in robotic surgery. In *International Workshop on Machine Learning in Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 556–573.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015.
24. Jha, K.K.; Dutta, H.S. Nucleus and cytoplasm-based segmentation and actor-critic neural network for acute lymphocytic leukaemia detection in single cell blood smear images. *Med. Biol. Eng. Comput.* **2020**, *58*, 171–186. [CrossRef] [PubMed]
25. Francia, G.A.; Pedraza, C.; Aceves, M.; Tovar-Arriaga, S. Chaining a U-net with a residual U-net for retinal blood vessels segmentation. *IEEE Access* **2020**, *8*, 38493–38500. [CrossRef]
26. Ding, H.; Cen, Q.; Si, X.; Pan, Z.; Chen, X. Automatic glottis segmentation for laryngeal endoscopic images based on U-Net. *Biomed. Signal Process. Control* **2022**, *accepted*. [CrossRef]
27. Siddique, N.; Sidike, P.; Elkin, C.; Devabhaktuni, V. U-Net and its variants for medical image segmentation: Theory and applications. *arXiv* **2020**, arXiv:2011.01118. [CrossRef]
28. Mur-Artal, R.; Tardós, J.D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans. Robot.* **2017**, *33*, 1255–1262. [CrossRef]
29. Iglovikov, V.; Shvets, A. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv* **2018**, arXiv:1801.05746.
30. Shvets, A.; Rakhlin, A.; Kalinin, A.A.; Iglovikov, V. Automatic instrument segmentation in robot-assisted surgery using deep learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17 December 2018.
31. Mahmoud, N.; Cirauqui, I.; Hostettler, A.; Doignon, C.; Soler, L.; Marescaux, J.; Montiel, J.M. ORBSLAM-based endoscope tracking and 3D reconstruction. In Proceedings of the International Workshop on Computer-Assisted and Robotic Endoscopy, Athens, Greece, 17 October 2016.
32. Mountney, P.; Stoyanov, D.; Yang, G.Z. Three-dimensional tissue deformation recovery and tracking. *IEEE Signal Process. Mag.* **2010**, *27*, 14–24. [CrossRef]
33. Piccinelli, N.; Roberti, A.; Tagliabue, E.; Setti, F.; Kronreif, G.; Muradore, R.; Fiorini, P. Rigid 3D registration of pre-operative information for semi-autonomous surgery. In Proceedings of the 2020 International Symposium on Medical Robotics (ISMR), Atlanta, GA, USA, 18–20 November 2020; pp. 139–145. [CrossRef]
34. 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.