applied sciences

MDPI

Article

Research on Surgical Gesture Recognition in Open Surgery Based on Fusion of R3D and Multi-Head Attention Mechanism

Yutao Men [1,2], Jian Luo [1,2,3], Zixian Zhao [1,2,3], Hang Wu [3], Guang Zhang [3], Feng Luo [3] and Ming Yu [3,*]

[1] Tianjin Key Laboratory for Advanced Mechatronic System Design and Intelligent Control, School of Mechanical Engineering, Tianjin University of Technology, Tianjin 300384, China
[2] National Demonstration Center for Experimental Mechanical and Electrical Engineering Education, Tianjin University of Technology, Tianjin 300384, China
[3] Systems Engineering Institute, Academy of Military Sciences, People's Liberation Army, Tianjin 300161, China
* Correspondence: yuming_1990@outlook.com

Abstract: Surgical gesture recognition is an important research direction in the field of computer-assisted intervention. Currently, research on surgical gesture recognition primarily focuses on robotic surgery, with a lack of studies in traditional surgery, particularly open surgery. Therefore, this study established a dataset simulating open surgery for research on surgical gesture recognition in the field of open surgery. With the assistance of professional surgeons, we defined a vocabulary of 10 surgical gestures based on suturing tasks in open procedures. In addition, this paper proposes a surgical gesture recognition method that integrates the R3D network with a multi-head attention mechanism (R3D-MHA). This method uses the R3D network to extract spatiotemporal features and combines it with the multi-head attention mechanism for relational learning of these features. The effectiveness of the R3D-MHA method in the field of open surgery gesture recognition was validated through two experiments: offline recognition and online recognition. The accuracy at the gesture instance level for offline recognition was 92.3%, and the frame accuracy for online recognition was 73.4%. Finally, its performance was further validated on the publicly available JIGSAWS dataset. Compared to other online recognition methods, the accuracy improved without using additional data. This work lays the foundation for research on surgical gesture recognition in open surgery and has significant applications in process monitoring, surgeon skill assessment and educational training for open surgeries.

Keywords: surgical gesture recognition; open surgery; R3D; multi-head attention mechanism; artificial intelligence; deep learning

check for updates

Citation: Men, Y.; Luo, J.; Zhao, Z.; Wu, H.; Zhang, G.; Luo, F.; Yu, M. Research on Surgical Gesture Recognition in Open Surgery Based on Fusion of R3D and Multi-Head Attention Mechanism. *Appl. Sci.* **2024**, *14*, 8021. https://doi.org/10.3390/app14178021

Academic Editor: José Salvador Sánchez Garreta

Received: 30 July 2024
Revised: 23 August 2024
Accepted: 4 September 2024
Published: 7 September 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Surgical gestures, as fine-grained surgical actions intentionally carried out by surgeons during surgery, carry precise control over tiny details, and their actions can produce explicit, perceptible and meaningful results [1]. Recognizing surgical gestures not only aids in understanding the surgeon's intentions and monitoring the state of the operation but also effectively detects and alerts to errors [2], thereby enhancing the safety and efficiency of the surgical process.

Currently, research on surgical gesture recognition has primarily focused on robotic surgery, while studies in the field of traditional surgery, particularly open surgery, are relatively scarce. This is partly because surgical robots can easily acquire digital video data and capture quantitative instrument motion trajectories that are difficult to obtain in traditional surgery. These data streams have facilitated the development of data-driven computational models in computer-assisted intervention (CAI) and surgical data science (SDS) [3]. On the other hand, there are many kinds of operations in open surgery, and the acquisition and annotation of video data also face great technical challenges and ethical restrictions. Despite the evident advantages of robotic surgery over traditional surgery, in

certain surgical scenarios, traditional surgery, especially open surgery, remains the standard treatment method, with broader indications than robotic surgery. In extreme situations (e.g., battlefield injuries and disasters), open surgery is still indispensable [4]. Therefore, this paper is dedicated to investigating fine-grained surgical gesture recognition in open surgery to explore the application value of this technology in the field of open surgery.

In open surgery, the research approach for surgical gesture recognition based on video data is similar to that of previous robotic surgery. Suppose a video of length T is given, containing a series of video frames v_t, $t = 1, \ldots, T$. In the task of surgical gesture recognition, predictions need to be made for each time step $t = 1, \ldots, T$ to determine the surgical gesture category $g(t) \in G$ at time t, where $G = \{1, \ldots, G\}$ is the set of surgical gestures. The methods of surgical gesture recognition vary based on the amount of information provided and can be categorized as follows: (i) using only the current video frame, i.e., $g(t) = g(v_t)$ (frame recognition); (ii) using the sequence of frames from the current time step to previous ones, i.e., $g(t) = g(v_k, \ldots, v_t)$, $k \geq 1$ (online recognition); or (iii) using the complete sequence of frames of the video, i.e., $g(t) = g(v_1, \ldots, v_T)$ (offline recognition) [5].

Previous research on surgical gesture recognition has mainly focused on robotic surgery, employing methods that include probabilistic graphical models and deep learning models. In the realm of probabilistic graphical models, hidden Markov models (HMMs) [6,7] and conditional random fields (CRFs) [8] have been widely used. Although these methods use interpretable features, they typically focus on a few adjacent frames, neglecting long-term temporal dependencies, and require dense kinematic annotations, which are not always available in surgery [9]. However, in recent years, the field has gradually shifted towards deep learning methods. For instance, recurrent neural networks (RNNs) [10] have been widely applied, with long short-term memory (LSTM) networks frequently used for long sequence modeling. Nonetheless, LSTMs have inherent gradient vanishing problems, thus limiting their ability to capture long-term video context [11]. Consequently, researchers have started to seek other methods better suited for handling long sequence data. Lea et al. proposed temporal convolutional networks (TCNs) [12], which use dilated convolutions to expand the receptive field of the convolution layers, capturing longer-range temporal dependencies. Although this model performs well in processing local adjacent information, it is generally less effective at capturing global dependencies. On the other hand, Funke et al. utilized 3D convolutions (3D-CNNs) [5] to extract spatiotemporal features from video segments and recognize and classify surgical gestures at the segment level. While 3D convolutions effectively extract spatiotemporal features, they lack the ability to model long-term dependencies in sequential data. Zhang et al. [13] proposed a symmetric dilated convolution method combined with an attention mechanism, achieving significant success in robotic gesture segmentation. However, this method's input and output are both complete surgical video information, limiting its application in online gesture recognition to retrospective detection post-surgery. To overcome these limitations, Gazis et al. [14] introduced the C3Dtrans network, which combines 3D convolution with a Transformer for online surgical gesture recognition. They employed the I3D model architecture as a feature extractor and used the Transformer architecture to model sequences of extracted deep features. However, due to the large number of parameters, this framework cannot be trained end-to-end.

Considering the current state of research, this paper proposes a method for recognizing open surgery gestures based on a combination of R3D [15] and multi-head attention mechanisms [16], named R3D-MHA. We use an unpre-trained R3D model to extract fine-grained surgical gesture features and combine it with a multi-head attention mechanism to capture the complex dependencies between long sequences, addressing the limitations of previous methods in capturing long-distance dependencies. Our method not only focuses on fine-grained operations but also captures long-distance dependencies. It also has some online recognition capabilities and allows for end-to-end training of the model. Due to the lack of datasets for surgical gesture recognition research in the field of open surgery, we constructed a database simulating open surgery and named it AEDCSSAD. It includes

three open surgical procedures: bowel anastomosis, hepatic rupture repair and closure of the abdomen surgeries. Currently, the research team has completed the definition and annotation of surgical gestures for the closure of the abdomen simulation surgery, which is a pioneering achievement.

Aiming at the problem of surgical gesture recognition in open surgery, we used the R3D-MHA model to conduct studies on both offline and online recognition tasks. The offline recognition task aims to classify manually extracted fine-grained gesture segments, while the online recognition task processes the input video stream to recognize surgical gestures in real time for each frame using a sliding window approach. Finally, we validated our method on the public dataset JIGSAWS to further test its ability to recognize robotic gestures.

In summary, our contributions are as follows:

- We constructed a simulated open surgery dataset and, with the assistance of professional surgeons, defined a surgical gesture vocabulary suitable for suture-based open surgeries, completing all annotations for the dataset.
- We proposed an open surgery gesture recognition method combining R3D and multi-head attention mechanisms. The applicability of this method for recognizing open surgery gestures was validated through offline and online recognition tasks.
- In addition to evaluating and analyzing the method on the open surgery dataset, we also tested the proposed model on the public JIGSAWS dataset for online recognition tasks and compared it with other online recognition models used for robotic gestures. This experiment demonstrated the potential application of our model in the field of surgical gesture recognition.

2. Materials and Methods

2.1. Dataset Overview

The self-constructed damage control surgery database of the research group provided a closure of the abdomen simulation surgery dataset for this study. This dataset was created by nine surgeons, including three practicing physicians, three resident interns and three clinical undergraduates. Each surgeon performed five simulated surgeries on a mannequin, resulting in a total of 45 video recordings. These data were collected using three different devices, as illustrated in Figure 1b. In the figure: 1 represents the surgical lamp coaxial camera, which captures third-person perspective surgical video data with a field of view aligned with the surgical lamp's light field; 2 represents the Kinect depth camera, positioned beside the operating table to capture third-person perspective video data; 3 represents a head-mounted camera worn by the operating surgeon, capturing first-person perspective video data as the view changes with the surgeon's head movements. The basic parameters of the equipment used for data collection are listed in Table 1. Due to the stable field of view provided by the surgical lamp coaxial camera—without the jitter issues of the head-mounted camera and the limited field of view of the Kinect camera—this study used video data captured by the surgical lamp coaxial camera. The videos have a frame rate of 30 fps, a resolution of 1200 $\times$ 720 and a duration ranging from 2 min 42 s to 5 min 22 s. All operations were performed on simulated organs, which were connected to a surgical manikin. The closure of the abdomen simulation surgery used a standard interrupted suture technique, requiring surgeons to use a needle holder to grasp the needle and suture the abdominal opening step by step. Figure 1 shows example video frames (a), the surgical scene (b) and the surgical manikin and simulated organs (c) from the closure of the abdomen simulation surgery.



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

Figure 1. Dataset overview (**a**), Sample video frames of closure of the abdomen surgery, (**b**) surgical scenes, number 1 is the Surgical lamp coaxial camera, number 2 is the Kinect depth camera, and number 3 is the Head-mounted video camera, (**c**) surgical manikin and simulated organs.

Table 1. Video acquisition equipment.

Device Name	Device Model	Data Format	Resolution	Frame Rate
Surgical lamp coaxial camera (Shanghai Pinxing Medical Equipment Co., LTD., Shanghai, China)	WYD2015-LC	Structured light video data *.mkv	1920 × 1080	30.0
Kinect depth camera (Microsoft, Redmond, WA, USA)	Kinect v2.0	Deep grayscale video data *.avi	512 × 424	25.0
Head-mounted video camera (GoPro, Inc., San Mateo, CA, USA)	HERO4 Black	Structured light video data *.MP4	1920 × 1080	25.0

2.2. Surgical Gesture Description

Previous researchers have compared surgical actions to human language, as surgical actions are also essentially a combination of basic activities performed sequentially under specific constraints [17]. By studying the language of surgical actions, scholars believe that techniques applied in the fields of human language and speech can be used to model surgical actions [1]. Therefore, detailed analysis and deconstruction of surgical actions have become particularly important, leading to the emergence of research and definition of fine-grained surgical gestures in this context. Given the lack of clearly defined standards for decomposing surgical gestures in open surgery, we referred to gesture definitions from related literature in the field of robotic surgery [1,18] and, with the assistance of professional surgeons, completed the definition of surgical gestures for the closure of the abdomen surgery. We defined the gestures from the perspective of a complete surgery, which differs from previous definitions focused on individual surgical activities.

In this study, we designed an open surgery gesture vocabulary containing 10 elements to describe the surgical gestures in closure of the abdomen surgery. As shown in Table 2, the gesture descriptions defined in this paper and their corresponding boundary frames are presented. These boundary frames mark the end of the surgical gestures. Based on these boundary frames, we performed frame-by-frame annotations on 45 video datasets of the temporary laparotomy surgery.

Table 2. Vocabulary of Surgical Gestures.

Gestures	Gesture Description	Boundary Frame
G1	Clamping the needle	Forceps just touching the tissue
G2	Forceps pick up the tissue	The holder has just started to move
G3	Positioning needle	The needle has just made contact with the tissue
G4	Pushing the needle through tissue	The needle tip has penetrated the tissue
G5	Clamping the needle through tissue	The needle has just departed from the tissue
G6	Pulling suture with holder	The holder loosens the needle or moves out of view
G7	Hand pulling suture	The needle holder approaches the vicinity of the suture line
G8	Knotting the suture	The scissors are not visible in the field of view
G9	Cutting the suture	The scissors are removed from the field of view
G10	Other	The holder regrasps the needle, Actions other than the nine surgical gestures (G1–G9), Other scenarios

As shown in Figure 2, for the distribution of surgical gestures in an example video, the following sequence of gestures is required for an ideal closure of the abdomen surgery: G1→G2→G3→G4→G5→G6→G7→G10→G8→G9.

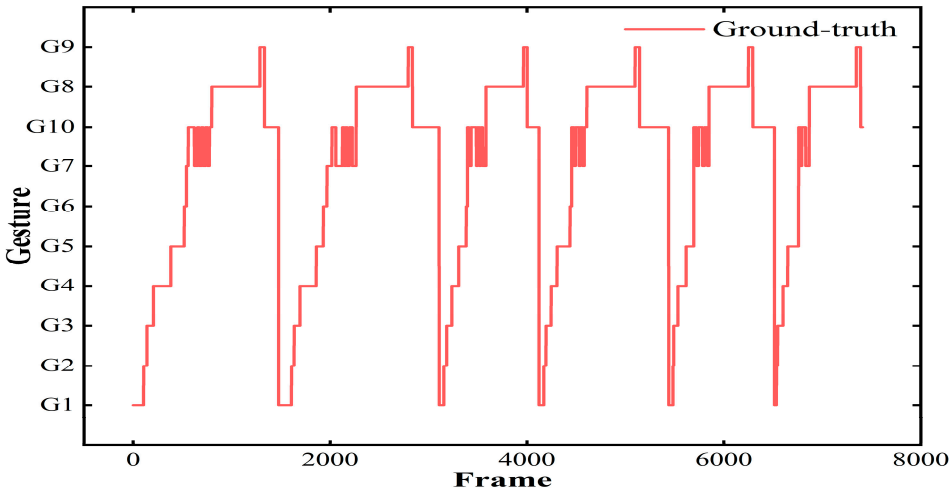


Figure 2. Distribution of surgical gestures in a sample video.

2.3. Data Processing

In this paper, the model is tested on two ways of recognition, hence the dataset was processed twice. First, we performed preprocessing by randomly selecting 3 complete videos from the 45 video datasets to evaluate the frame-wise accuracy of online recognition. For the remaining 42 videos, we segmented them according to predefined boundary frames and manually extracted action segments of each gesture category, known as gesture instances. Subsequently, we divided the gesture instances of these 10 gesture categories into training, validation and test sets in a ratio of 0.64:0.16:0.2. For each gesture instance, we uniformly sampled at least 32 frames. If the number of frames for a gesture instance was less than 32, we opted to duplicate the last frame to fill the sequence to 32 frames, rather than discarding these instances. Ultimately, the numbers of instances in the training, validation and test sets were 1948, 491 and 615, respectively.

For the offline recognition task, we processed the gesture instances into image sequences with a height H = 171, width W = 128 and sequence length T = 32 as input to the model. For the online recognition task, we further processed the gesture instances to obtain fine-grained segment representations. To ensure the model accurately recognizes gesture transitions, we applied overlapping processing to the extracted 32-frame gesture instances using a sliding window approach with a step size of 2 and sequence length of 16, i.e., moving 2 frames backward each time. There was an overlap of 14 frames between adjacent sequences. Finally, the numbers of overlapping segments obtained for the training,

validation and test sets were 31,708, 8496 and 10,103, respectively. Table 3 shows the data related to surgical gestures in the dataset.

Table 3. Data related to surgical gestures.

Gesture	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
Instances	168	163	220	218	219	203	326	218	217	665
Overlapping segments	1495	1347	2376	4144	3086	1490	2493	17,173	2857	13,846
Mean frame count	33	28	45	108	75	18	23	432	59	110
Min frame count	3	5	6	21	11	2	3	99	26	3
Max frame count	314	97	170	356	297	73	135	1435	216	593

2.4. R3D-MHA Network Architecture

The proposed R3D-MHA model has a modular structure, composed of the R3D module and the multi-head attention mechanism module. The first part involves using the R3D module as the backbone network to extract spatiotemporal features from video sequences. The R3D model is a 3D-CNN extended from the ResNet architecture, with a structure similar to 2D ResNet but replacing 2D convolutions with 3D convolutions, and using a $1 \times 3 \times 3$ convolutional kernel in the temporal dimension. As shown in Figure 3, the main structure of the R3D network includes an STConv layer, four STResLayer layers, and an STPool layer. The STConv layer uses a standard 3D convolution module to extract features with a kernel size of $3 \times 7 \times 7$. The STResLayer is composed of a basic residual module, as depicted in Figure 3, which consists of two 3D convolutional layers, each followed by a BatchNorm layer and a Softplus layer. Additionally, the STPool layer uses AdaptiveAvgPool3d to dynamically adjust the pooling layer size. Notably, in this study, all activation functions in R3D were replaced with Softplus instead of ReLU, enabling more effective weight updates and improving the model's generalization capability.

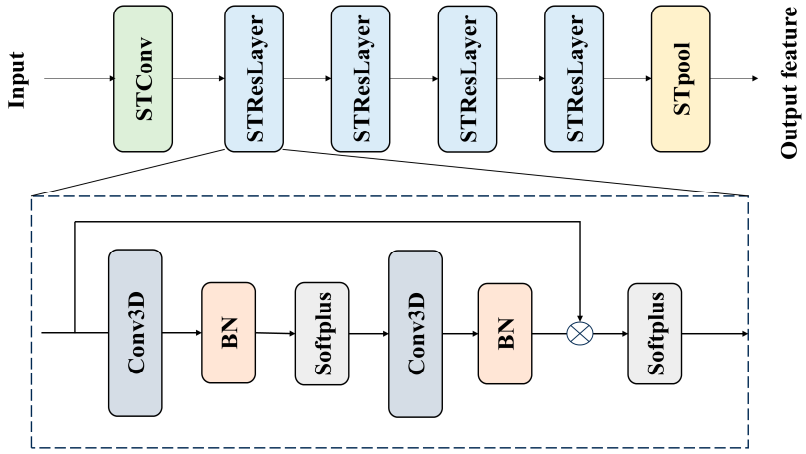


Figure 3. R3D model structure.

The second part uses a multi-head attention mechanism module to further process the extracted features. It first performs three linear transformations on the input feature data, mapping the input (X) into matrices representing the query (Q), key (K) and value (V), respectively. Then, according to the following Equation (1), it calculates the attention weights for each head. $Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V$, where ($d_k$) is the dimension of the attention head, and (W_i^Q, W_i^K, W_i^V) are the weight matrices learned by this head. Finally, multiple attention heads are used to compute attention weights in parallel to obtain multiple output results. These outputs are then linearly transformed and concatenated as

shown in Equation (2) to obtain the final multi-head attention representation. Figure 4 is the schematic diagram of the multi-head attention mechanism.

$$head_i = Attention(Q_i, K_i, V_i) = softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (1)$$

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_h) W^O \quad (2)$$

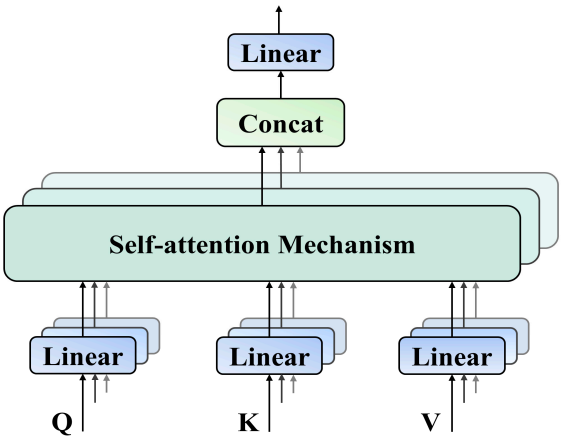


Figure 4. Schematic diagram of multi-head attention mechanism.

In summary, the two modules that constitute the R3D-MHA model are responsible for extracting spatiotemporal features from video sequences and performing global processing. By integrating these methods, the model can fully leverage the strengths of both modules, enhancing its capability to capture and process features from video sequences.

2.5. Overall Technical Framework

Figure 5 illustrates the overall technical framework of this paper. We conducted two different experiments where the R3D-MHA model had two different inputs. For offline recognition, we performed evaluations at the gesture level by inputting the processed 32-frame (T = 32) images into the R3D model to generate 512-dimensional deep features. Subsequently, the multi-head attention module conducted relevance learning on the deep features, and ultimately generated gesture labels G_i at the gesture level using Softmax. For online recognition, we input the processed 16-frame overlapping segments, but this paper does not focus on the results of the segment-level test set. We trained the model by inputting overlapping segments and then evaluated the frame-wise accuracy of online recognition on three complete videos.

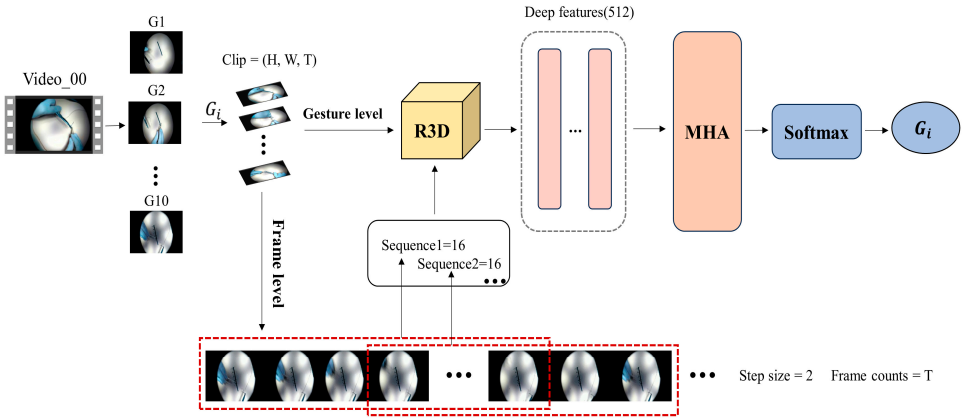


Figure 5. Overall technical framework diagram.

2.6. Implementation Details

Each set of experiments was assigned the same random seed number to ensure reproducibility of the results. All models were fine-tuned on our self-built dataset without using pre-trained models. All model codes in the experiments were written based on the PyTorch framework, and the experiments were run in an RTX 3090 (24219MiB) GPU environment. All experiments were set to 70 training epochs, except for the Vivit model which was trained for 100 epochs. The batch size was set to 16, the optimizer used was SGD, the learning rate was set to 10^{-3} and the loss function was the cross-entropy function.

2.7. Evaluation Metrics

To quantitatively analyze the performance of the seven networks trained in this study, we employed four different evaluation metrics. In the offline recognition task, we used three common classification metrics in the field of action recognition, namely accuracy (Equation (3)), precision (Equation (4)) and recall (Equation (5)). Among these, TP represents true positives, TN represents true negatives, FP represents false positives and FN represents false negatives. We calculated the precision and recall for each gesture category and then averaged these values across all categories to obtain the corresponding values for the entire test set. Accuracy indicates the average accuracy of samples across the entire test set. In the online recognition task, this paper uses frame-wise accuracy (Equation (6)). Here, TC denotes the number of correctly predicted frames, and TF represents the total number of frames in a complete video. The subtraction of 16 accounts for the input sequence length of 16 frames.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Frame - wiseaccuracy = \frac{TC}{TF - 16} \tag{6}$$

3. Results

3.1. Offline Recognition

In the offline recognition task, we trained seven models on the open surgical dataset, including the baseline model R3D and six other models, including three classic 3D convolutional models: R(2+1)D [15], I3D [19] and C3D [20], the dual-path model SlowFast [21] and the Transformer-based model Vivit [22] applied in the field of video understanding. Figure 6 illustrates the training process of these seven models, where (a) shows the training loss and (b) displays the results of the test set every 10 epochs. From the figure, it can be observed that at the 40th epoch, the loss value reaches its minimum and stabilizes, and the test set accuracy also reaches a stable level.

Table 4 presents the final test results of the models. It can be seen that our proposed R3D-MHA model surpasses the baseline R3D model by 1.9%, 2.0% and 2.0% in terms of accuracy, precision, and recall, respectively. Additionally, it leads in all metrics compared to the other five models.

Given the imbalance in the sample sizes of our self-built dataset, this study also considers the model's recognition performance for individual gesture categories. We use the recall rate of single gesture recognition to evaluate the recognition accuracy of various surgical gestures. As shown in Table 5, overall, the gestures G1, G2 and G7 are easily confused with other surgical gestures, leading to lower recognition accuracy. This may be because the sample sizes of G1 and G2 are relatively small, and G7 (hand pulling suture) is somewhat similar to G10 (other actions), both of which result in the hand being out of the field of view. Although the R3D-MHA model shows a slight decrease in the accuracy of recognizing G4 and G5 compared to the R3D model, with a reduction of 0.02, it improves

the recognition accuracy of the three easily confused gestures G1, G2 and G7 by 0.02, 0.03 and 0.09, respectively. Furthermore, it shows varying degrees of improvement for other gestures. This also demonstrates the effectiveness of combining the R3D model with the multi-head attention mechanism.

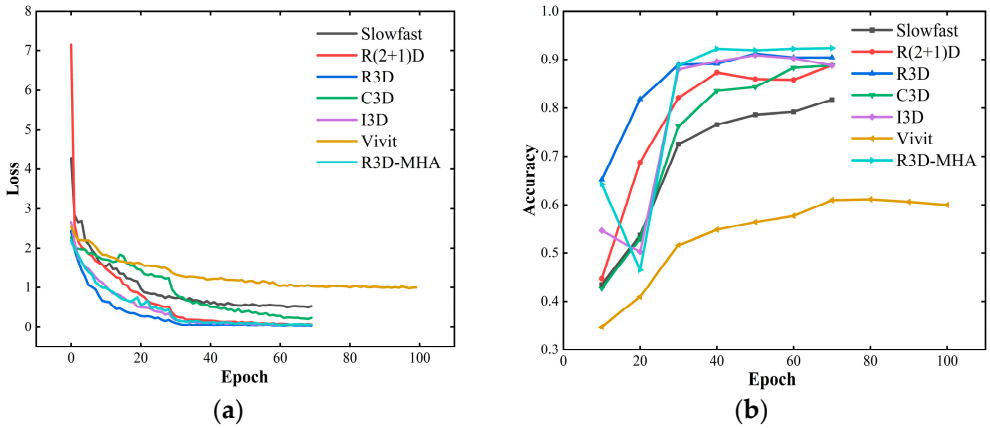


Figure 6. Training process. training set loss (**a**), test set accuracy (**b**).

Table 4. Experimental results. The best results and methods by column-wise are shown in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)
R(2+1)D	88.7	89.0	87.8
C3D	88.6	88.0	87.2
Slowfast	81.7	79.7	78.3
Vivit	60.0	56.3	55.6
I3D	88.8	88.6	88.1
R3D	90.4	90.5	90.0
R3D-MHA	**92.3**	**92.5**	**92.0**

Table 5. Recall of each model for each gesture instance. The best results and methods by column-wise are shown in bold.

Model	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
R(2+1)D	0.85	0.74	0.81	0.92	0.94	0.94	0.78	0.90	0.94	0.94
C3D	0.59	0.67	0.91	0.93	0.95	0.73	0.76	0.86	0.91	0.92
Slowfast	0.73	0.72	0.53	0.81	0.89	0.83	0.77	0.60	0.90	0.93
Vivit	0.73	0.44	0.60	0.48	0.32	0.45	0.70	0.44	0.62	0.79
I3D	0.85	0.80	0.94	0.92	0.90	0.87	0.81	0.85	0.94	0.92
R3D	0.88	0.79	0.92	**0.96**	**0.96**	0.89	0.76	0.92	0.98	0.93
R3D-MHA	**0.90**	**0.82**	**0.96**	0.94	0.94	**0.94**	**0.85**	**0.92**	**0.98**	**0.94**

3.2. Online Recognition

For the online recognition task, we performed online inference using the R3D-MHA model on three reserved full-length videos. We processed the videos with a sliding window of 16 frames and a step size of 1 frame, meaning the recognition of the *n*-th frame was based on the information from frames [n − 15, n]. No future frames were provided to the model during the experiment to ensure this method's applicability to online gesture recognition, performing real-time gesture recognition based solely on existing video information. Table 6 presents the frame-wise accuracy (FWA) and average results of the three full-length videos. We compared the online recognition results of the R3D model and found that the R3D-MHA model improved the frame accuracy by 0.4% in the online recognition task. Additionally, we presented visualized recognition results of the R3D-MHA model on the best- and worst-performing videos, as shown in Figure 7. The blue line represents the predicted values, and the red line represents the ground truth. The more the prediction overlaps with the ground

truth, the higher the prediction accuracy. From the figure, it can be seen that although there are some error fluctuations in recognizing transitions of the doctor's actions, since we trained a classification model rather than a segmentation model, the R3D-MHA model can still accurately identify the doctor's gesture transitions, demonstrating its potential for online recognition applications.

Table 6. Online recognition results of three complete videos. The best results and methods by column-wise are shown in bold.

Model	Video00	Video01	Video02	Average (%)
R3D	61.3	75.7	82.1	73.0
R3D-MHA	62.6	75.7	81.8	73.4

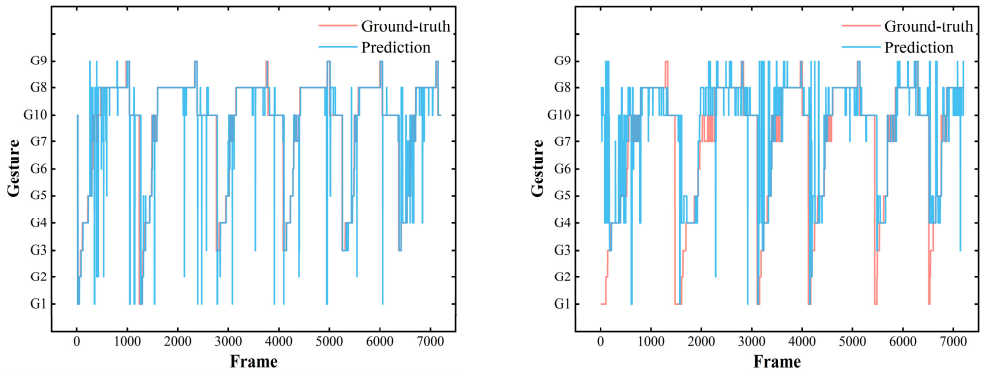


Figure 7. Comparison of Best and Worst visualization results. Best (**a**), Worst (**b**).

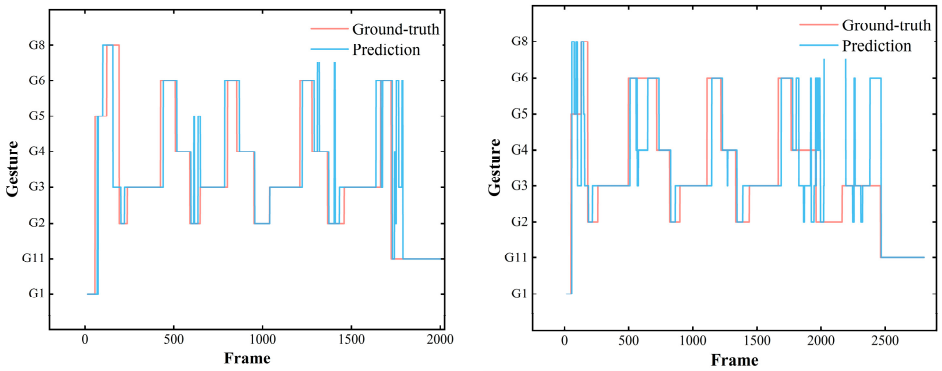
3.3. Evaluation of the JIGSAWS Dataset

In addition to our open surgical dataset, this paper also evaluates the performance of the R3D-MHA model on the publicly available JIGSAWS dataset in the field of robotic gesture recognition. The JIGSAWS dataset includes three basic surgical tasks: suturing (SU), knot tying (KT) and needle passing (NP), and provides two types of data: tool kinematics and RGB videos. The dataset also offers two cross-validation schemes: Leave-One-Supertrial-Out (LOSO) and Leave-One-User-Out (LOUO). In this study, we used 39 RGB video samples from the suturing task to evaluate the performance of the R3D-MHA model. This task was performed by eight participants, each conducting five experiments.

In this study, to maintain consistency with previous online recognition validation methods, we used the Leave-One-User-Out (LOUO) validation scheme. Specifically, five full video data from one participant were selected to test the model's frame-wise accuracy for online recognition, while the remaining video data were processed according to the aforementioned method, with the step size for extracting overlapping segments adjusted to 1 due to the relatively short length of these videos. The specific experimental details are consistent with those described earlier. Table 7 shows the results of the LOUO three-fold cross-validation, with the best fold achieving 77.9% and the average of the three folds being 76.3%. We reported the frame-wise accuracy (FWA) for each complete video and the average results. Figure 8 shows the visualized recognition results for the best and worst videos.

Table 7. Online recognition results of 5 complete videos on JIGSAWS.

Fold	SU-001	SU-002	SU-003	SU-004	SU-005	Average (%)
1	77.9	83.7	74.4	69.3	67.1	74.5
2	66.6	80.6	79.5	79.9	82.9	77.9
3	67.2	X	80.7	76.3	81.7	76.5
All						76.3



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

Figure 8. Comparison of Best and Worst visualization results on JIGSAWS, Best (**a**), Worst (**b**).

Table 8 shows the comparison of the accuracy of the proposed R3D-MHA method with some previous online surgical gesture recognition methods on the SU task of JIGSAWS. The results show that the R3D-MHA method outperforms methods based on manually extracted features and are 0.5% higher than those the C3Dtrans method, which uses complete video streams for online rolling input. It is worth noting that although some models that perform better than R3D-MHA, such as 3D-CNN, MTL-VF and CNN+LC-SC-CRF, used additional training data, our model still achieved good results without using any pre-trained models or additional data. Furthermore, in our experiments, the validation method used was to test on complete video data, which also demonstrates the effectiveness of the R3D-MHA model in practical applications.

Table 8. Accuracy comparison for surgical gesture recognition on the SU task of JIGSAWS. The best results and methods by column-wise are shown in bold.

Model	Acc (%)	Trained on Additional Dataset	Applicable Online
CRF (dense) [6]	68.8	-	√
MsM-CRF (STIP–STIP) [6]	66.3	-	√
MsM-CRF (dense–dense) [6]	71.8	-	√
CNN+LC-SC-CRF [23]	76.6	√ (Sensor Values)	√
ST-GCN [24]	67.9	-	√
MTL-VF [25]	82.1	√ (Sports-1M and ImageNet)	√
3D-CNN [5]	84.0	√ (Kinetics)	√
C3Dtrans [14]	75.8	-	√
R3D-MHA	**76.3**	-	√

4. Discussion

Addressing the current lack of research on open surgery gesture recognition, this study draws on previous surgical gesture recognition research in robotic surgery. We established a closure of the abdomen simulation dataset to study surgical gesture recognition in open surgery. Additionally, this paper proposes a surgical gesture recognition method based on open surgery video data, named R3D-MHA. This method uses a modular structure, combining the R3D network with the multi-head attention mechanism, achieving end-to-end training and inference. Experimental results show that in experiments on the open

surgery dataset, this method performs excellently in both offline and online recognition tasks, achieving a gesture instance level accuracy of 92.3% for offline recognition and a frame-wise accuracy of 73.4% for online recognition. Furthermore, we conducted a more comprehensive evaluation of the R3D-MHA model on the public JIGSAWS dataset, using the Leave-One-User-Out (LOUO) cross-validation method. The results show that without using any pre-trained models or additional data, the recognition accuracy of this method improves compared to other online recognition methods.

Of course, this study still has some limitations. Firstly, the improvement in results for the R3D-MHA method in the online recognition task is limited, as we have only conducted preliminary exploration on online detection of open surgery gestures without further optimization of the online recognition algorithm. Secondly, the dataset is based on simulated surgical scenarios, which differ from real surgical environments in certain aspects, such as video noise and blood interference, potentially impacting the model's generalization ability in real-world settings. Additionally, although our dataset includes data from surgeons with three different skill levels, which makes the model have certain generalization ability, it may also limit the exploration of variations in surgical skills among different surgeons.

In summary, the contribution of this study lies in the construction of a dataset for open surgery gesture recognition research and the proposal of the R3D-MHA model for both offline and online gesture recognition in open surgery. We have thoroughly validated its performance in recognizing surgical gestures in open surgery. In future research, we plan to further improve the database. On one hand, we will expand it to include more types of open surgeries, such as bowel anastomosis and hepatic rupture repair. On the other hand, we will collect data that more closely resemble real-world environments to enhance the applicability and accuracy of the research. Additionally, we will explore the application of surgical gestures in actual surgeries, such as technical skill assessment based on surgical gestures and surgical workflow monitoring, to improve the model's suitability for intelligent operating rooms. We will also investigate solutions to challenges like the lack of standardized segmentation for surgical gestures and difficulties in data annotation.

Author Contributions: Conceptualization, Y.M.; methodology, J.L. and M.Y.; software: F.L.; validation, H.W., M.Y. and G.Z.; formal analysis, H.W., M.Y. and G.Z.; investigation, J.L.; resources, H.W. and M.Y.; data curation, J.L. and Z.Z.; writing—original draft preparation, J.L.; writing—review and editing, H.W., M.Y. and G.Z.; visualization, J.L.; supervision, Y.M and M.Y.; project administration, M.Y.; funding acquisition, M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available upon reasonable request from the corresponding author.

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

References

1. Gao, Y.; Vedula, S.S.; Reiley, C.E.; Ahmidi, N.; Varadarajan, B.; Lin, H.C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D.D.; et al. JHU-ISI gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In Proceedings of the Modeling and Monitoring of Computer Assisted Interventions (M2CAI)—MICCAI Workshop, Boston, MA, USA, 25 September 2014; Volume 1, pp. 1–10.
2. Yasar, M.S.; Alemzadeh, H. Real-time context-aware detection of unsafe events in robot-assisted surgery. In Proceedings of the 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), IEEE, Valencia, Spain, 29 June–2 July 2020.
3. Maier-Hein, L.; Vedula, S.S.; Speidel, S.; Navab, N.; Kikinis, R.; Park, A.; Eisenmann, M.; Feussner, H.; Forestier, G.; Giannarou, S.; et al. Surgical data science for next-generation interventions. *Nat. Biomed. Eng.* **2017**, *1*, 691–696. [CrossRef] [PubMed]
4. Zhao, Z.; Gu, J. Open surgery in the era of minimally invasive surgery. *Chin. J. Cancer Res.* **2022**, *34*, 63–65. [CrossRef] [PubMed]

5. Funke, I.; Bodenstedt, S.; Oehme, F.; von Bechtolsheim, F.; Weitz, J. Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019, Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019*; Springer International Publishing: Cham, Switzerland, 2019.
6. Tao, L.; Zappella, L.; Hager, G.D.; Vidal, R. Surgical gesture segmentation and recognition. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013, Proceedings of the 16th International Conference, Nagoya, Japan, 22–26 September 2013*; Part III; Springer: Berlin/Heidelberg, Germany, 2013.
7. Lea, C.; Hager, G.D.; Vidal, R. An improved model for segmentation and recognition of fine-grained activities with application to surgical training tasks. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, IEEE, Waikoloa, HI, USA, 5–9 January 2015.
8. Mavroudi, E.; Bhaskara, D.; Sefati, S.; Ali, H.; Vidal, R. End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Lake Tahoe, NV, USA, 12–15 March 2018.
9. Zhang, J.; Nie, Y.; Lyu, Y.; Yang, X.; Chang, J.; Zhang, J.J. SD-Net: Joint surgical gesture recognition and skill assessment. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 1675–1682. [CrossRef] [PubMed]
10. DiPietro, R.; Lea, C.; Malpani, A.; Ahmidi, N.; Vedula, S.S.; Lee, G.I.; Lee, M.R. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016, Proceedings of the 19th International Conference, Athens, Greece, 17–21 October 2016*; Part I; Springer International Publishing: Berlin/Heidelberg, Germany, 2016.
11. Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. *PMLR* **2013**, *28*, 1310–1318.
12. Lea, C.; Vidal, R.; Reiter, A. Temporal convolutional networks: A unified approach to action segmentation. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016; Part III. Springer International Publishing: Berlin/Heidelberg, Germany, 2016.
13. Zhang, J.; Nie, Y.; Lyu, Y.; Li, H.; Chang, J.; Yang, X. Symmetric dilated convolution for surgical gesture recognition. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020, Proceedings of the 23rd International Conference, Lima, Peru, 4–8 October 2020*; Part III; Springer International Publishing: Berlin/Heidelberg, Germany, 2020.
14. Gazis, A.; Karaiskos, P.; Loukas, C. Surgical gesture recognition in laparoscopic tasks based on the transformer network and self-supervised learning. *Bioengineering* **2022**, *9*, 737. [CrossRef] [PubMed]
15. Tran, D.; Wang, H.; Torresani, L.; Ray, J.; Lecun, Y.; Paluri, M. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 30, Proceedings of the Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017*; MIT Press: Cambridge, MA, USA, 2017.
17. Lin, H.C. Structure in Surgical Motion. Ph.D. Thesis, Johns Hopkins University, Baltimore, MD, USA, 2010.
18. Ahmidi, N.; Tao, L.; Sefati, S.; Gao, Y.; Lea, C.; Haro, B.B.; Zappella, L.; Khudanpur, S.; Vidal, R.; Hager, G.D. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Trans. Biomed. Eng.* **2017**, *64*, 2025–2041. [CrossRef] [PubMed]
19. Carreira, J.; Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
20. Tran, D.; Bourdev, L.D.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015.
21. Feichtenhofer, C.; Fan, H.; Malik, J.; He, K. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
22. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
23. Rupprecht, C.; Lea, C.; Tombari, F.; Navab, N.; Hager, G.D. Sensor substitution for video-based action recognition. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Daejeon, Republic of Korea, 9–14 October 2016.
24. DiPietro, R.; Ahmidi, N.; Malpani, A.; Waldram, M.; Lee, G.I.; Lee, M.R.; Vedula, S.S.; Hager, G.D. Segmenting and classifying activities in robot-assisted surgery with recurrent neural networks. *Int. J. Comput. Assist. Radiol. Surg.* **2019**, *14*, 2005–2020. [CrossRef] [PubMed]
25. Wang, T.; Wang, Y.; Li, M. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Lima, Peru, 4–8 October 2020; Volume 1, pp. 668–678.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.