IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 22, 2020, accepted August 10, 2020, date of publication August 18, 2020, date of current version September 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3017523

Multi-Granular Semantic Analysis Based on Nasal Endoscopic Video

XIAOYING PAN[1,2], HAO ZHAO[1], NI LIU[1], AND HONGYU WANG[1,2]

[1]School of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China
[2]Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

Corresponding author: Hongyu Wang (hywang@xupt.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFC0121502, in part by the Innovation Fund of Xi'an University of Posts and Telecommunications under Grant CXJJLZ2019041 and Grant CXJJLY2019097, in part by the Scientific Research Project of Education Department of Shaanxi Provincial Government under Grant 19JK0808, and in part by the Special Fund of the Key Laboratory of Network Data Analysis and Intelligent Processing of Shaanxi Province.

ABSTRACT The semantic analysis of nasal endoscopic video is a challenging task since lots of irrelevant and insignificant information exists in the untrimmed surgical video, i.e. background, blur, judder or blood-stained video fragments. It is important to identify the start and end point of the valid surgical fragments automatically and remove the invalid fragments of endoscopic surgery videos for medical education & research. However, the performance of deep-learning based methods, which use a fixed time interval and a sliding window, are severely affected when the interference information appears randomly in the nasal endoscopic video. Specifically, the surgical video is a continuous process globally, while many local discontinuity fragments are brought when endoscope enters and exits the cavity frequently. Hence, we propose a multi-granularity semantic analysis framework that can simultaneously meet the accuracy and timeliness required for endoscopic surgery video semantic analysis. Our approach is an end-to-end solution. First, a joint model is created to extract the temporal-spatial features of the surgical video on a coarse-grained scale. Meanwhile, an attention mechanism is used to automatically select the informative spatial features of endoscopic video. Second, a hierarchical self-correction module is proposed to correct the boundaries of the surgical operation iteratively on a fine-grained scale. Finally, we justify the proposed network through extensive experiments and quantitative comparisons against other state-of-the-art approaches. We achieve a good performance in terms of accuracy and efficiency.

INDEX TERMS Multi-granular hierarchical, nasal endoscopic surgery, self-correction, video semantic analysis.

I. INTRODUCTION

Endoscopic surgery has been more and more practiced in nasal surgery in recent years because of its less trauma and quick recover [1]–[3], the number of nasal surgery videos was continuously booming. These videos provided a great basis for documentation, training of young surgeons [4], medical research [5] and analytics in healthcare [6].

Usually, a complete endoscopic surgical video is recorded from the beginning of the operation to the end of the operation. Not only the surgical operation fragments are preserved, but also some unrelated surgical operations such as covering the endoscope lens with blood stains, defocusing the lens during movement, and cleaning the endoscope lens are also retained. However, doctors only need the valid video clips after the surgery. They have to edit the video to make it more convenient. It is not only difficult and time consuming for Doctors to manually edit the video but also is very expensive to ask a third-party agency, such as SurgiCast (https://www.surgicast.io/medical-video-editing), to edit [7]. There is a great opportunity for researcher to develop the methods to automate the editing of endoscopic surgery videos. Semantic analysis of endoscopic surgical videos is one of the most important keys in the automation [8]. As is shown in Figure1. Semantic analysis methods not only are able to analyze the start and end point of the surgical operation, but also can analyze the invalid images in the operation. Endoscopic surgery video is characterized by continuity and discontinuity. Continuous surgical operations are interrupted by these invalid shots in the endoscopic surgery video.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

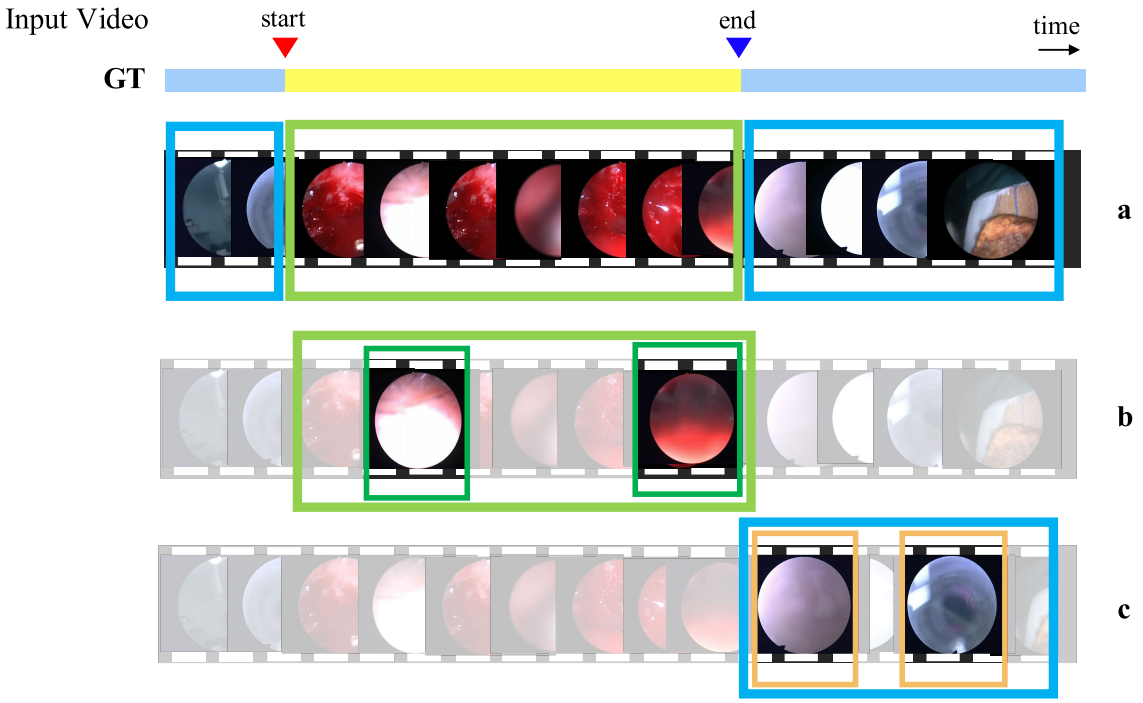


FIGURE 1. An overview of our task. (a) An untrimmed video of an endoscopic surgery. We need to determine the start and end points of each surgical procedure segment. (b) Analyze the blurred image of the surgical procedure. (c) Analysis of the blurred image of the endoscope outside the cavity.

Especially, the randomness of discontinuity block the way to find the start point and end point. Moreover, the operation of the surgery is a continuous process, but a complete operation is split into discrete pieces due to various phenomena such as the need to clean the endoscope lens. And these interruptions are random, there is no regularity at all.

Most of the researches were focused on the lesion detection [9], lesion segmentation [10], and lesion diagnosis [11], all of which were performed on a single frame of image. On the other hand, there were studies on classification of gynecological organs, eight kinds of surgical operation recognition customized in abdominal surgery video [12]. However, there were relatively few studies on semantic analysis of nasal surgery videos [13]. Popular methods usually used a fixed time interval [14] or a sliding window [15] to generate candidate proposals and perform semantic analysis in the field of natural scene video. But these methods were not very effective in semantic analysis of endoscopic surgery video because of the random discontinuity of endoscopic surgery videos.

In this article, we propose a new framework for semantic analysis on endoscopic surgery videos via a deep neural network, which is called Multi-granular Hierarchical Network (MHN) as is shown in Figure2. First, a four classification was performed on successive n key frames by using an end-to-end spatial-temporal feature modeling. After obtaining a preliminary prediction sequence result, a more granular correction was applied for a hierarchical self-correction module. Finally, the automatic marking of the surgical operation was implemented, and the automatic editing was completed. From inputting the original video into the network and outputting the edited effective surgical screen video, the whole process did not require human participation. It was a fully automatic processing mode.

In summary, the key contributions of our work include:

- This work provides the first semantic analysis for nasal endoscopic surgery video using deep learning method. And we propose a framework that automatically detects non-surgical operations in endoscopic surgery video.
- Semantic analysis of endoscopic surgery video with multi-granular spatial-temporal features combined with modeling scheme.
- The hierarchical structure of the self-correction module from rough to fine is proposed to improve the accuracy of surgical video semantic analysis.
- Compared with state-of-art performance [16], [17], our method further improves the accuracy on our dataset to 89%.

The rest of this article is organized as follows. In Section 2 some relevant works are reviewed. In Section 3, we describe the details of our proposed approach. In Section 4, we present the experiments and results. Finally, we present our concluding remarks in Section 5.

II. RELATED WORK

A. ENDOSCOPIC IMAGE PROCESSING

Recently, Deep convolutional neural network (CNN) [18]–[20] has made breakthroughs in various task such as

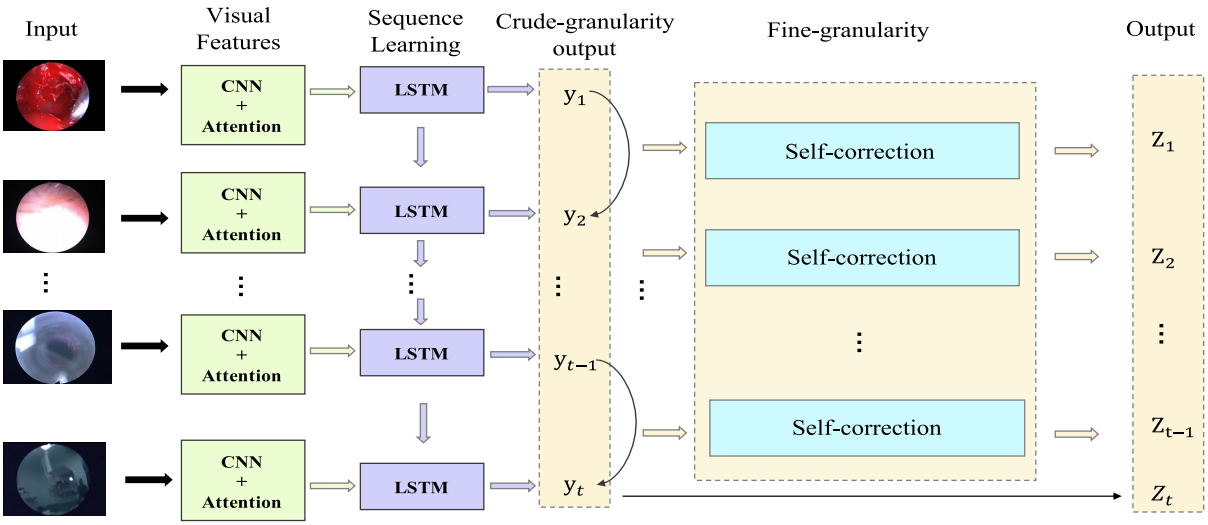


FIGURE 2. An overview of the semantic analysis framework. We first extract keyframes from the video and input successive t keyframes into a CNN that incorporates the attention module. Then, the feature map extracted from the CNN is input into the LSTM module for sequential learning. Finally, the results of the previous step are entered into a hierarchical self-correction module for more precise semantic analysis.

image classification [18], image segmentation [19], object detection [20] and so on. Despite the difference between natural and medical images, deep learning has been imported from endoscopic image processing and presented impressive performance on polyp recognition [21], bleeding detection [22], and polyp classification [23]. At the same time, deep learning has also made great progress of solving some specific problems of the field of medical imaging. For example, deep learning is used to study deformable registration methods of medical images [24]. Detection of respiratory diseases from medical images of heuristic algorithms [25]. Bacterial recognition model composed of regional covariance of convolutional neural network [26]. And Ibtehaz and Rahman [27] used MultiResUnet network to segment multi-peak medical images. Further, the semantic analysis of surgical video based on deep learning technology has gradually gained the attention of researchers [13]. For example, Twinanda *et al.* [28] used CNN to extract image features from laparoscopic cholecystectomy video, and migrated the pre-trained Alexnet model to the medical field, in combination with the hidden Markov model. Finally, a single frame image recognition rate of 92.2% was obtained. Petscharnig and Schöffmann [12] used CNN and support vector machine models to identify eight surgical operations that were customized in the video of abdominal surgery. These works are based on the single frame image of the surgical video. Although CNN effectively improves the ability to express features, the process of processing a video stream into a single frame tends to ignore hidden features in nasal endoscopic surgery videos, which makes it difficult to improve the accuracy of nasal endoscopic surgery video analysis.

B. VIDEO SEMANTIC ANALYSIS

Although few researches on the semantic analysis methods of medical videos were developed, there were many new methods in the natural scene video. Action recognition and temporal action detection are two important branches of video semantic analysis and has been extensively studied [29]–[32].

Action recognition models can be used to extract summary-level visual features in untrimmed video. Action recognition has been extensively studied in the past few years [29]–[33]. Earlier methods are mostly based on hand-crafted visual features such as HOF, HOG and MBH [33]. In recent years, two-stream network [29], [30], [34]and C3D network [31], [32], [35] learns appearance and motion features. Typically, two-stream network learns appearance and motion features based on RGB frame and optical flow field separately. For example, Lin *et al.* [30] proposed a Boundary Sensitive Network (BSN), which used two sub-networks (spatial network and temporal network) for encoding video information. Because this kind of method modeled the spatiotemporal features of video separately, it was easy to ignore the relevance. The defects of this method are gradually exposed in many tasks. C3D network adopt 3D convolutional layers to capture appearance and motion features directly from the original frame. For example, Xu *et al.* [32] introduced a spatial-temporal feature-preserving filter in a C3D network to maximize the resolution of the video in the time dimension, which improved the accuracy of video frame-by-frame recognition effectively. However, the 3D network has the higher requirements on data and hardware, and the training difficulty had to be improved. On the other hand, the method often performs poorly for the scenes with frequent video

shot switching. The 3D convolutional network did not improve the performance of video content parsing tasks significantly although it overcame the shortcomings of the above Two Stream convolutional network.

Temporal action detection task aimed to detect action instances in untrimmed videos including temporal boundaries and action classes, and could be divided into proposal and classification stages. Earlier works [36] directly used sliding windows for the proposal generation. Recently some methods [14], [37] generated the proposals with pre-defined temporal durations and intervals, and used multiple methods to evaluate the confidence score of proposals, such as dictionary learning [14] and recurrent neural network [37]. These two methods had a good semantic analysis effect in natural videos with continuous features, especially standard predefined actions. However, these methods for semantic analysis may have some major disadvantages due to the discontinuity of endoscopic video: (1) usually not temporally precise,and surgical video requires more precise positioning; (2) Fixed pre-defined temporal durations and intervals are not suitable for randomly occurring invalid images.

At the same time, the Visual Question Answering (VQA) task is also a new field that involves action detection, but VQA not only needs to focus on action detection, but also needs to understand the text. And VQA also works on single-frame images. So its method is not applied to our work.

Compared to these methods, our multi-granularity semantic analysis method is superior to in two aspects: (1) Coarse-grained analysis overcomes the random discontinuity of endoscopic surgery video. (2) Fine-grained hierarchical self-correction more accurately locates the boundaries of surgical operations.

III. MULTI-GRANULAR SEMANTIC ANALYSIS

The semantic analysis of surgical video has become more difficult because of the coexistence of video continuity and discontinuity in endoscopic surgery. We propose a spatial-temporal combined framework MHN to solve this problem as is shown in Figure2. Generally speaking, from input to output, no human intervention is required. After inputting the original video, through the processing of the model, the output only retains meaningful video clips. Firstly, a coarse-grained semantic analysis is performed on the combination of spatial-temporal features. The coarse-grained analysis combines the spatial and temporal characteristics of CNN and RNN networks, and introduces the attention mechanisms on the spatial network to enhance the learning of spatial features. Thereby ensuring the accuracy of the analysis and taking into account the timeliness of the surgical video analysis. Secondly, the hierarchical correction of coarse-grained results provide more precise positioning of surgical action boundaries based on the timing relationship.

A. DATA DEFINITION

An untrimmed video is a sequence of frames. The Key frame or I-frame was defined as a single frame of digital

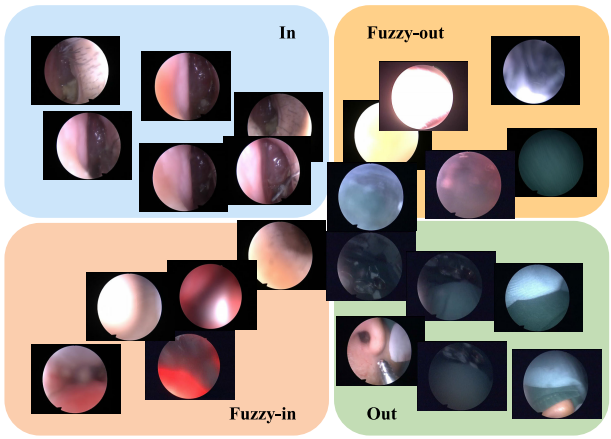


FIGURE 3. Examples of nasal endoscopic images in our dataset. (In) Clear operation shots inside the nasal cavity. (Out) Clear shots outside the nasal cavity. (Fuzzy-in) Blurred shots inside the nasal cavity. (Fuzzy-out) Blurred shots outside the nasal cavity.

content that the compressor examines independent of the frames that precede and follow it and stores all of the data needed. The video sequence can be denoted as $X = \{x_n\}_{n=1}^{k}$ where x_n is the nth key frame in X. Key frame. In our work, we extract a key frame every fifteen original frames and mark the time points and labels for each key frame. The nasal endoscopic image was pre-defined as four labels Y = In, Out, Fuzzy-in, Fuzzy-out} by the medical professional. As is shown in Figure3, Our data samples have large internal differences and small differences between categories, which will make semantic analysis difficult. In particular, there are big differences not only between the In category and the Fuzzy-In category but also between the Out category and the Fuzzy-Out category. In general, after semantic analysis of the nasal surgery video, only the In category shots are kept during editing. However, sometimes in order to maintain the continuity of the surgical video, the Fuzzy-In category shots are usually retained.

B. CRUDE-GRANULARITY ANALYSIS ON SPATIAL-TEMPORAL FEATURES

As is shown in Figure4, after extracting the key frames of the surgical video, CNN was used to learn spatial features. An attention mechanism was introduced to perform feature tracking based on the particularity of the endoscopic image. Further, the time characteristics were learned through the Recurrent Neural Network (RNN) network, a coarse-grained sequence was generated as the result.

We applied a deep neural network architecture, ResNet-50 developed by He *et al.* [38]. as our spatial feature extractor. ResNet is a deep CNN architecture containing residual learning blocks to address a problem of degradation during learning very deep networks. The output of each block of ResNet-50 has half spatial resolution compared to that of the previous block. Various settings for the feature extractor have been tested, including deeper ResNet-101, different designs

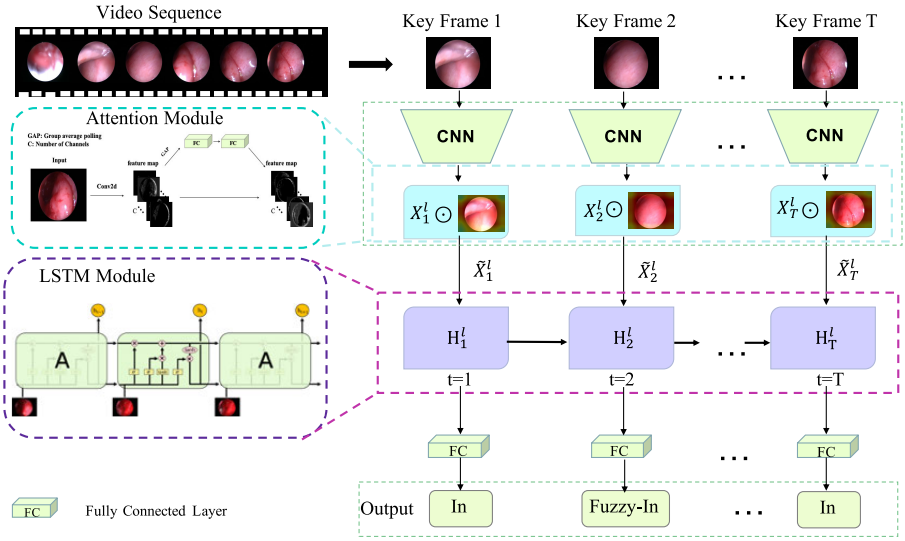


FIGURE 4. An overview of the Crude-granularity framework. The video sequence is decomposed into key frames and entered into the CNN model with the fully connected layer removed. The Attention module is introduced into the CNN model. After obtaining the result sequence predicted by the spatial model, the LSTM module is added to perform the time dimension modeling.

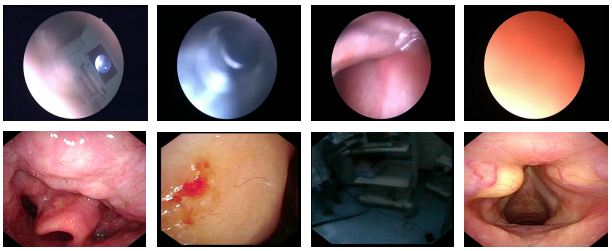


FIGURE 5. Examples of endoscopic images imaged by different endoscopic devices.

of the convolution neural network, and up sampling to the image width 512. The results are similar. Therefore, the simpler and computationally efficient setting was chosen.

The endoscopic image has a distinct difference from other images. Since the shapes of endoscopes with different specifications are different, the final image area is an irregular polygon or a circle. As is shown in the Figure5, the shape of the endoscope lens of different manufacturers is also different. There are many studies on the analysis of irregularly shaped endoscopic images. For the related work of detection in the circular area of endoscopic video [39], the MultiResUNet [27] network is used to solve the problems of different scales of medical images. In our work, we introduced the SENet [40] attention module on the spatial network to track the effective information of the image. SENet can automatically obtain the importance of each feature channel through learning, and then use this importance to enhance useful features and suppress features that are not very useful for the current task. As is shown in the Figure4, through the CNN network convolution transformation, a two-dimensional feature map with a channel number of C and a feature map size of H*W was obtained. It is input into the attention module unit as an input feature. First, in the spatial dimension, through the global average pooling layer, each two-dimensional feature channel will be transformed into a real number. This real number was used to characterize the global receptive field of the feature map, and the output dimension is consistent with the input feature channel number. Then, a Bottleneck structure was formed by two fully connected layers to model the correlation between channels, and the same number of weights as the input features were output. First, the feature dimension was reduced to 1/16 of the input, and then activated by ReLU and then returned to the original dimension through a Fully Connected layer. Compared with using a Fully Connected layer directly, it had more nonlinearities and could better fit the complex correlation between channels. It also greatly reduced the amounts of parameters and calculations. Then we used a Sigmoid activation function to obtain the normalized weight between 0 and 1, and finally we used the scale operation to weight the normalized weight to the characteristics of each channel. As is shown in Eq 1, s refers to the weight sequence output by the attention module,σ refers to the ReLU activation function,δ refers to the sigmoid activation function,$W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$, $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$, and z refers to the real number obtained by the global average pooling layer.

$$s = \sigma(W_2 \delta(W_1 z)). \tag{1}$$

Compared with some methods for detecting circular areas, the use of attention mechanism is more universal. And we avoid the method of dividing the image first and extracting the effective area before processing.

Since our job is to perform semantic analysis on endoscopic video, the key frames extracted from the original video not only have spatial features, but at the same time, the temporal characteristics of continuous key frames are

also what we want to learn. It is known the key frame can be judged roughly by the preceding and succeeding frames when the video is continuous. We use RNN to capture global information and long-term dependencies, which is able to learn patterns and long-term dependencies from sequential data. Moreover, LSTM [41] is a type of RNN architecture that stores information about its predictions in other regions of the cell state, it can predict the classification of key frames based on the relationship of consecutive frames. LSTM has the characteristics of selective memory that can control the transmission status through the gated state: Remembering the useful information for a long time and ignoring the unnecessary information. As is shown in Eq2, z^f performs the forgetting control that keeps the previous memory cells that should be retained and be forgotten, z^i is the selects memory control that selects important information to record.

$$c^t = z^f \cdot c^{t-1} + z^i \cdot z. \tag{2}$$

We used the CNN network to do local feature extraction, and used the LSTM network to model the timing relationship of consecutive frames. The combination of the two networks could simultaneously took into account the spatial characteristics of the key frame image and the temporal characteristics of consecutive key frames. As is shown in Figure 4, we used the LSTM module to model continuous key frames in order to obtain more temporal feature information in the endoscopic video. We added an LSTM module behind the CNN network based on the attention module. Each training needs to input n consecutive key frames and output n classification results. In our work, we first trained the resnet-50 model with the attention module added. On this basis, we removed the last layer of the fully connected layer and fine-tune it to obtain the 512-dimensional features of the output and used it as an input to connect a unidirectional LSTM network. The LSTM network had 512 neurons and 5 times step. Therefore, the input of the CNN network was a vector unit composed of 5 consecutive key frames. After the LSTM module, the predicted key frame category was output through a fully connected layer. We set 4 neurons for the fully connected layer to correspond to the four key frame categories.

C. FINE-GRAINED SEMANTIC ANALYSIS ON TIME SERIES RELATIONSHIP

In coarse-grained analysis, it is not sufficiently accurate to use keyframes as sample sequences for surgical operation boundary localization. In order to make the boundaries of the cropped effective video clips more precise, we proposed a fine-grained hierarchical self-correction module to solve this problem.

In the coarse-grained module, the CNN model is used to analyze the key frame image sequence and obtain the preliminary result sequence. In addition, in the result sequence, two adjacent key frames with different types of results are found. In this way, it can be considered that the previous frame is the end point of the previous candidate video, and the next frame is the start point of the next candidate video. At a coarse

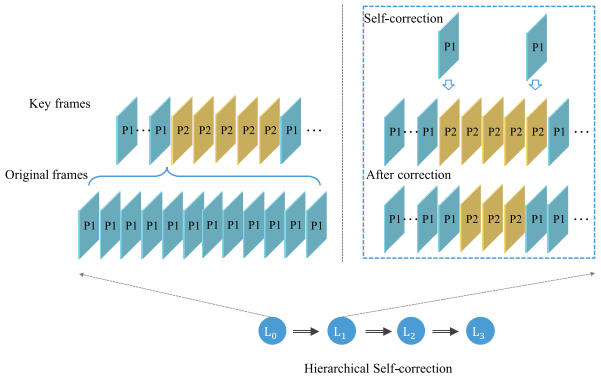


FIGURE 6. Schematic diagram of layered self-correction. (a) From the key frames of adjacent candidate video clips, extract the original frames between these two key frames to reconfirm the boundary frame results. (b) Update the key frame sequence by the self-correction result obtained from the original frame.

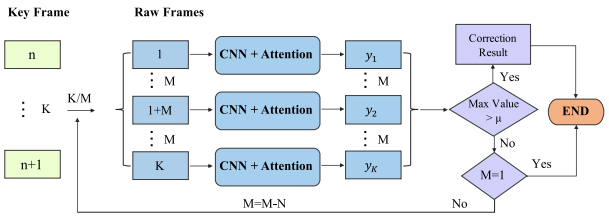


FIGURE 7. Hierarchical Self-correction Flowchart. K: All original frames in the middle of consecutive keyframes. M: sampling interval. y_i: Result label. Max Value: Number of largest categories. N: Update interval sampling parameters. μ: Threshold.

granularity, in order to consider the timeliness and accuracy of video analysis, we use key frames to analyze the video. In order to improve the accuracy of the video clip boundaries, we performed a hierarchical analysis of the original frames at the beginning and end of the candidate video. As is shown in Figure 6 (a), for adjacent P1 and P2, the original frame in the middle of these two key frames is extracted to reconfirm the boundary frame result. Figure 6 (b) assumes that after the original frame discrimination, the judgment results of the start and end points of the P2 candidate segment have changed. Update the key frame sequence with the correction result. The updated original frame sequence will obtain new candidate fragments. In the fine-grained layered correction, we have designed a total of 3 layers of correction. The example in the figure where L_0 points to L_1 refers to the first layer of correction. The specific calibration process is shown in Figure 7. We assume that there are K original frames between the nth key frame and the n + 1th key frame. Each layer samples K/M frames at intervals of M frames. For these sampled original frames, we analyze and calculate the results through the spatial feature model. If the number of categories in the sampled original frame is the largest and greater than the threshold μ, then we regard the category as a valid result, use this result to update the nth key frame result, and end the analytic hierarchy process. Otherwise, by updating the interval

TABLE 1. Data set distribution.

	In	Out	Fuzzy-In	Fuzzy-Out	Total
Training	4216	4028	1411	1632	11287
Validation	1550	1547	849	1357	5303
Test	678	287	73	155	1193
Total	6444	5862	2333	3144	17783

sampling, N is subtracted from M to update to the new M value, thereby sampling more original frames and continuing the analysis. Finally, when M is updated to 1, the maximum number of categories does not exceed the threshold, we end the fine-grained module and retain the original results.

IV. EXPERIMENT

A. EXPERIMENTAL SETTINGS

1) DATASETS

Our dataset is based on the key frame of 2 hours of clinical nasal endoscopic surgery video extraction from the First Affiliated Hospital of Xi'an Jiaotong University,China. There are total 17783 images. These images are marked in the four shot categories. The number of each categories are displayed in the Table1. Among them, the training set, verification set and test set are randomly distributed according to a ratio of about 7: 2: 1. These surgical videos are from 12 different nasal surgeries. The model of the endoscope is Olympus ENF-T3 with a resolution of 720 * 576. And the preprocessing is 224 * 224 when inputting the model.

2) IMPLEMENTATION DETAILS

Optimization was performed using synchronous SGD with momentum 0.9, a learning rate of 0.001 and decay of 0.0001. The entire experiment was implemented using Python 3.5, based on the Tensorflow 1.18.0 environment, running on two 12 GB Nvidia Tesla K80 GPU machine with batch size 16 for 100 epochs.

3) EVALUATION METRICS

the performance of the model quantitatively is measured by using four commonly used metrics where TP, TN, FP and FN denote the number of true-positive, true-negative, false-positive and false-negative detection results, respectively. Recall reflects the classification model's ability to identify positive samples. The higher the recall, the stronger the model's ability to identify positive samples. Precision reflects the model's ability to distinguish negative samples. The higher precision indicates the model's ability to distinguish negative samples. The higher the F1-score is, the more robust the classification model is.

B. RESULTS ANALYSIS

1) ANALYSIS OF COARSE-GRAINED RESULTS

First, we analyzed the effect of the coarse-grained model. Figure8 is a Class Activation Map (CAM) comparison chart between the coarse-grained model and backbone. From the CAM chart, the heat in the effective area is significantly

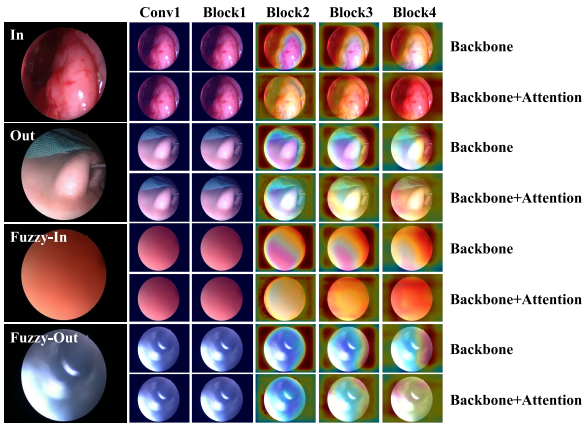


FIGURE 8. Through the class activation graph of each category, the effects of the first convolution layer and each residual block after adding the attention module are compared. The first column is the original image of each category.

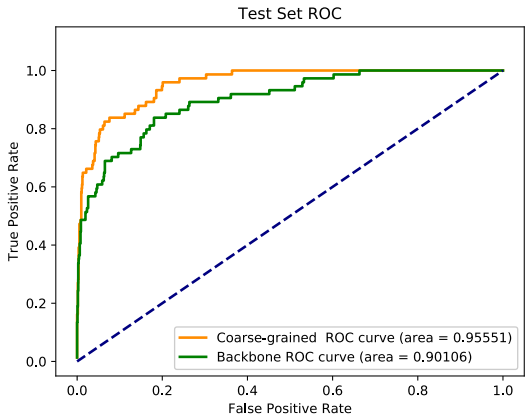


FIGURE 9. Example of ROC chart comparison between coarse-grained model and backbone model, orange is coarse-grained model, green is backbone model.

increased after the attention module is added to the backbone. More attention will be given on valid areas. Secondly, we compared and analyzed the ROC curves of the backbone and the coarse-grained model. The results shown that the effect of the coarse-grained model was significantly better than that of the backbone in Figure9.

In addition, we compared the effect of using only the backbone and adding the LSTM module to the backbone. The results were shown in Figure 9. After adding the LSTM module backbone, the accuracy rate reached 0.82, the weighted average accuracy, the recall rate and f1score reached 0.82. Compared with the backbone, the accuracy was improved by 7%. At the same time, we also compared the evaluation results of each analogy. Among them, the improvement of Out category and Fuzzy-In category was more obvious than the Fuzzy-out category. The analysis of the above results showed that for time-series tasks, CNN networks performed semantic analysis by learning spatial features and obtained

TABLE 2. The results of backbone model, coarse-grained model and MHN are compared under precision, recall and f1 score in macro avg and weighted avg. Also compares the feature types of accuracy and different models.

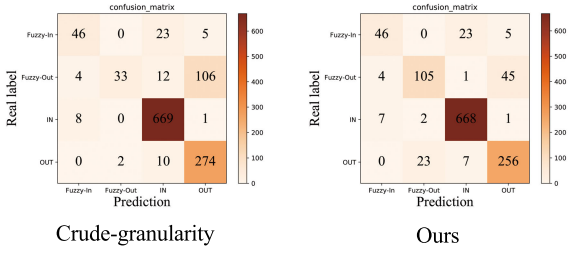
	feature type		Macro avg			Weighted avg			Accuracy
method	spatial	temporal	precision	recall	f1score	precision	recall	f1score	
backbone [38]	Yes	No	0.63	0.67	0.57	0.81	0.75	0.74	0.75
backbone+LSTM	Yes	Yes	0.71	0.70	0.70	0.82	0.82	0.82	0.82
Crude-granularity	Yes	Yes	**0.85**	0.69	0.71	0.87	0.86	0.83	0.85
Ours	Yes	Yes	0.84	**0.78**	**0.80**	**0.89**	**0.89**	**0.89**	**0.89**

TABLE 3. The backbone model, coarse-grained model, and MHN are compared with the precision, recall, and f1score of each category.

	In			Out			Fuzzy-In			Fuzzy-Out		
method	precision	recall	f1score	precision	recall	f1score	precision	recall	f1score	precision	recall	f1score
backbone [38]	**0.99**	0.80	0.88	0.64	**0.96**	0.76	0.32	**0.82**	0.47	0.59	0.11	0.18
backbone+LSTM	0.97	0.97	0.97	0.70	0.75	0.72	0.72	0.68	0.70	0.43	0.40	0.42
Crude-granularity	0.94	**0.99**	0.96	0.71	**0.96**	0.82	0.79	0.62	0.70	**0.94**	0.21	0.35
Ours	0.95	0.98	**0.97**	**0.81**	0.90	**0.85**	**0.80**	0.61	**0.79**	0.81	**0.61**	**0.70**

TABLE 4. The evaluation time of MHN and some state-of-the-art models in accuracy, total parameters, and semantic analysis of a surgical video of about 12 minutes were compared.

	Accuracy	Tota params	Times
ResNet [38]	0.7476	22,680,004	47.40s
InceptionV3 [42]	0.7694	21,810,980	86.02s
Xception [17]	0.8155	20,869,676	76.97s
Ours	0.8927	23,556,548	76.55s



Crude-granularity Ours

FIGURE 10. Confusion matrix of Crude-granularity and Ours.

good results. On this basis, through the LSTM module to further learn the timing characteristics, the performance of classification will be better.

Finally, as is shown in Table 2, compared with the backbone, the coarse-grained model has been partially improved. The accuracy rate reached 0.85, an increase of 10%. As is shown in Table 3, the results of the Fuzzy-In category show that the accuracy of the coarse-grained model reaches 0.79, which is 47% higher than the backbone accuracy, while the f1score reaches 0.70, which is an increase of 23%. The results of the Fuzzy-Out category show that the accuracy of the coarse-grained model reaches 0.94, 44% higher than the backbone, the recall rate is 10% higher than the backbone, and the f1score is 17% higher than the backbone. Compared with the network where the LSTM module is added to the backbone, the results of the coarse-grained model can prove the effect of the attention module. Increased accuracy by 3%. Precision increased by 5%, recall increased by 4%, and f1score increased by 1%. However, the effect is still not satisfied by analyzing the two index recall and f1score. Then we added the fine-grained modules to MHN and shown the results as follows.

2) FINE-GRAINED RESULTS ANALYSIS

Further, we analyzed the results of the fine-grained model from Table2, we concluded that the effects of the macro average and weighted average of our method were better than the backbone and coarse granularity. Especially, the recall and f1score are significantly improved, which mean that our method had better stability. MHN's weighted average precision, recall, and f1score all reached 0.89. Compared with coarse-grained, precision increased by 2%, recall increased by 3%, and f1score increased by 6%. Moreover, the performance of accuracy is 14% higher than backbone and 4% higher than the coarse granularity. The comparison results of each category shown in Table3 suggest that MHN has good classification accuracy and good stability in each category. In particular, recall in the Fuzzy-out category are 40% higher than coarse-grained, and f1score is 35% higher. As is shown in Figure 10, the left picture is the confusion matrix of the coarse-grained model, and the right picture is the confusion matrix of the fine-grained model. It can be clearly found that the fine-grained model improves the fuzzy-in category and the out category. In addition, we analyzed the effect of fine-grained model correction. The success of fine-grained model correction means that the starting and ending points of an effective surgical video will be more accurate, and the video viewing effect after editing will be smoother. Correction failure refers to a situation where the judgment error of the coarse-grained model cannot be corrected. According to the statistics of the test data, a total of 97 corrections were completed, of which 69 corrections were successful and the correction success rate was 71%.

Finally, as is shown in Table4, we compared the accuracy, total model parameters, and model processing time with several state-of-the-art indicators. Our model had the significantly higher accuracy than other models that only

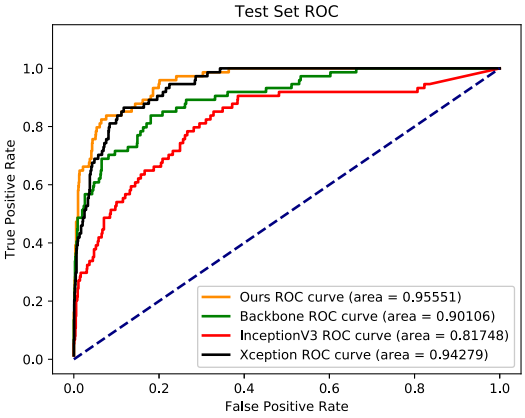


FIGURE 11. The ROC curves and AUC results of MHN and some state-of-the-art models are compared.

learned spatial features. The final accuracy rate reached 0.8927, which was an 8% improvement over Xception. The total parameter amount was not increased remarkably and the processing speed was also within an acceptable range. The processing time was shorter than of InceptionV3 and Xception although our module had more parameters. The main reason is that more time is cost in image pre-processing for both InceptionV3 and Xception.

As is shown in Figure11, we also compared the ROC with these methods. We could find that the ROC chart clearly reflected the performance advantage of our method for video analysis of nasal endoscopic surgery.

These results suggest the effectiveness of MHN. And MHN achieves the salient performance since it can generate proposals with (1) the attention module pays more attention to the effective image area in irregular endoscopic images. (2) The LSTM models continuous key frames to better capture the timing information in the endoscope video. (3) The self-correction module further accurately judges the boundaries of the surgical operation.

V. CONCLUSION

In this article, we present a framework for nasal endoscopic video semantic analysis. Our method can accurately and efficiently analyze the surgical operation part of the nasal endoscopic surgery video and remove the blurred frame. In experiments, we demonstrate that feature learning combined with spatial and temporal is better than spatial learning alone. Moreover, the hierarchical self-correction from coarse to fine further improves the accuracy of semantic analysis for nasal endoscopic video, and this hierarchical structure greatly improves the efficiency.

REFERENCES

[1] M. T. Myaing, D. J. MacDonald, and X. Li, "Fiber-optic scanning two-photon fluorescence endoscope," *Opt. Lett.*, vol. 31, no. 8, pp. 1076–1078, 2006. [Online]. Available: http://ol.osa.org/abstract.cfm?URI=ol-31-8-1076

[2] J. Peng, X. Li, H. Tang, L. Ma, T. Zhang, Y. Li, and S. Chen, "Miniaturized high-resolution integrated 360° electronic radial ultrasound endoscope for digestive tract imaging," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 5, pp. 975–983, May 2019.

[3] Y. Otsuka, K. Kamata, M. Takenaka, K. Minaga, H. Tanaka, and M. Kudo, "Electronic hydraulic lithotripsy by antegrade digital cholangioscopy through endoscopic ultrasound-guided hepaticojejunostomy," *Endoscopy*, vol. 49, no. 12, pp. 316–318, Dec. 2017.

[4] T. Jimbo, S. Ieiri, S. Obata, M. Uemura, R. Souzaki, N. Matsuoka, T. Katayama, K. Masumoto, M. Hashizume, and T. Taguchi, "Effectiveness of short-term endoscopic surgical skill training for young pediatric surgeons: A validation study using the laparoscopic fundoplication simulator," *Pediatric Surg. Int.*, vol. 31, no. 10, pp. 963–969, 2015, doi: 10.1007/s00383-015-3776-y.

[5] H. Lee, Y. Lee, C. Song, H. R. Cho, R. Ghaffari, T. K. Choi, K. H. Kim, Y. B. Lee, D. Ling, H. Lee, S. J. Yu, S. H. Choi, T. Hyeon, and D.-H. Kim, "An endoscope with integrated transparent bioelectronics and theranostic nanoparticles for colon cancer treatment," *Nature Commun.*, vol. 6, no. 1, p. 10059, Dec. 2015, doi: 10.1038/ncomms10059.

[6] B. Mánzer, K. Schoeffmann, and L. Böszörmenyi, "Content-based processing and analysis of endoscopic images and videos: A survey," *Multimedia Tools Appl.*, vol. 77, no. 1, pp. 1323–1362, Jan. 2018, doi: 10.1007/s11042-016-4219-z.

[7] M. J. Primus, K. Schoeffmann, and L. Boszormenyi, "Temporal segmentation of laparoscopic videos into surgical phases," in *Proc. 14th Int. Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2016, pp. 1–6.

[8] C. Loukas, "Video content analysis of surgical procedures," *Surg. Endoscopy*, vol. 32, no. 2, pp. 553–568, Feb. 2018.

[9] T. Falk, "U-net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019, doi: 10.1038/s41592-018-0261-2.

[10] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[11] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[12] S. Petscharnig and K. Schöffmann, "Learning Laparoscopic video shot classification for gynecological surgery," *Multimedia Tools Appl.*, vol. 77, no. 7, pp. 8061–8079, Apr. 2018, doi: 10.1007/s11042-017-4699-5.

[13] A. P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "EndoNet: A deep architecture for recognition tasks on Laparoscopic videos," *IEEE Trans. Med. Imag.*, vol. 36, no. 1, pp. 86–97, Jan. 2017.

[14] F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–9.

[15] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," *THUMOS14 Action Recognit. Challenge*, vol. 1, no. 2, p. 2, 2014.

[16] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

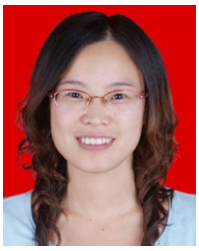
[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[19] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[21] Y. Yuan and M. Q.-H. Meng, "Deep learning for polyp recognition in wireless capsule endoscopy images," *Med. Phys.*, vol. 44, no. 4, pp. 1379–1389, Apr. 2017, doi: 10.1002/mp.12147.

[22] X. Jia and M. Q.-H. Meng, "Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and CNN features," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 3154–3157.
[23] R. Zhang, Y. Zheng, T. W. C. Mak, R. Yu, S. H. Wong, J. Y. W. Lau, and C. C. Y. Poon, "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 41–47, Jan. 2017.
[24] L. Mansilla, D. H. Milone, and E. Ferrante, "Learning deformable registration of medical images with anatomical constraints," *Neural Netw.*, vol. 124, pp. 269–279, Apr. 2020.
[25] M. Woániak and D. Poáap, "Bio-inspired methods modeled for respiratory disease detection from medical images," *Swarm Evol. Comput.*, vol. 41, pp. 69–96, Aug. 2018.
[26] D. Polap and M. Wozniak, "Bacteria shape classification by the use of region covariance and convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–7.
[27] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.
[28] A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to predict remaining surgery duration from laparoscopic videos without manual annotations," *IEEE Trans. Med. Imag.*, vol. 38, no. 4, pp. 1069–1078, Apr. 2019.
[29] S. Paul, S. Roy, and A. K. Roy-Chowdhury, "W-talc: Weakly-supervised temporal activity localization and classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 1–9.
[30] S. Wang, Z. Miao, W. Xu, C. Ma, and M. Li, "Boundary sensitive and category sensitive network for temporal action proposal generation," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2019, pp. 5194–5199.
[31] Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C.-W. Fu, and P.-A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Trans. Med. Imag.*, vol. 37, no. 5, pp. 1114–1126, May 2018.
[32] Y. Xu, C. Zhang, Z. Cheng, J. Xie, Y. Niu, S. Pu, and F. Wu, "Segregated temporal assembly recurrent networks for weakly supervised multiple action detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9070–9078.
[33] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
[34] W. Ye, J. Cheng, F. Yang, and Y. Xu, "Two-stream convolutional network for improving activity recognition using convolutional long short-term memory networks," *IEEE Access*, vol. 7, pp. 67772–67780, 2019.
[35] J. Zhang and H. Hu, "Deep spatiotemporal relation learning with 3D multi-level dense fusion for video action recognition," *IEEE Access*, vol. 7, pp. 15222–15229, 2019.
[36] S. Karaman, L. Seidenari, and A. Del Bimbo, "Fast saliency based pooling of Fisher encoded dense trajectories," in *Proc. ECCV*, 2014, vol. 1, no. 2, p. 5.
[37] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 768–784.
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[39] B. Munzer, K. Schoeffmann, and L. Boszormenyi, "Detection of circular content area in endoscopic videos," in *Proc. 26th IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2013, pp. 534–536.
[40] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.



XIAOYING PAN received the B.S. degree in computer communications from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2002, and the M.S. degree in computer application technology and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, in 2009.

Since 2019, she has been working as a Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications. She is the author of two books and more than 70 articles. Her research interests include medical image processing, artificial intelligence, and data mining.

HAO ZHAO received the B.S. degree in communication engineering from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2016, where he is currently pursuing the M.S. degree in computer technology. His research interests include medical image processing and deep learning.

NI LIU received the B.S. degree in electronic science and technology from the Mingde College, Northwestern Polytechnical University, Xi'an, China, in 2018. She is currently pursuing the M.S. degree in computer technology with the Xi'an University of Posts and Telecommunications, Xi'an. Her research interest includes temporal action detection in video analysis.

HONGYU WANG received the B.S. degree from Hebei United University, in 2012, and the Ph.D. degree from Northwest University, in 2018. She is currently a Lecturer with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China. Her main research interests include medical image processing, AI, and pattern recognition.