


Article

Hybrid Spatiotemporal Contrastive Representation Learning for Content-Based Surgical Video Retrieval

Vidit Kumar ¹, Vikas Tripathi ¹, Bhaskar Pant ¹, Sultan S. Alshamrani ², Ankur Dumka ³, Anita Gehlot ⁴, Rajesh Singh ⁴, Mamoon Rashid ^{5,*}, Abdullah Alshehri ⁶ and Ahmed Saeed AlGhamdi ⁷

- ¹ Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun 248002, India; viditkumaruit@gmail.com (V.K.); vikastripathi.be@gmail.com (V.T.); pantbhaskar2@gmail.com (B.P.)
- ² Department of Information Technology, College of Computer and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia; susamash@tu.edu.sa
- ³ Department of Computer Science and Engineering, Women's Institute of Technology, Dehradun 248007, India; ankurumka2@gmail.com
- ⁴ Division of Research & Innovation, Uttarakhand University, Dehradun 248007, India; anita.ri@uttarakhanduniversity.ac.in (A.G.); rajeshsingh@uttarakhanduniversity.ac.in (R.S.)
- ⁵ Department of Computer Engineering, Faculty of Science and Technology, Vishwakarma University, Pune 411048, India
- ⁶ Department of Information Technology, Al Baha University, P.O. Box 1988, Al Baha 65731, Saudi Arabia; aashehri@bu.edu.sa
- ⁷ Department of Computer Engineering, College of Computer and Information Technology, Taif University, P.O. Box 11099, Taif 21994, Saudi Arabia; asjannah@tu.edu.sa
- * Correspondence: mamoon.rashid@vupune.ac.in; Tel.: +91-7814346505



Citation: Kumar, V.; Tripathi, V.; Pant, B.; Alshamrani, S.S.; Dumka, A.; Gehlot, A.; Singh, R.; Rashid, M.; Alshehri, A.; AlGhamdi, A.S. Hybrid Spatiotemporal Contrastive Representation Learning for Content-Based Surgical Video Retrieval. *Electronics* **2022**, *11*, 1353. <https://doi.org/10.3390/electronics11091353>

Academic Editor: Leonardo Galteri

Received: 10 March 2022

Accepted: 20 April 2022

Published: 24 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 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/>).

Abstract: In the medical field, due to their economic and clinical benefits, there is a growing interest in minimally invasive surgeries and microscopic surgeries. These types of surgeries are often recorded during operations, and these recordings have become a key resource for education, patient disease analysis, surgical error analysis, and surgical skill assessment. However, manual searching in this collection of long-term surgical videos is an extremely labor-intensive and long-term task, requiring an effective content-based video analysis system. In this regard, previous methods for surgical video retrieval are based on handcrafted features which do not represent the video effectively. On the other hand, deep learning-based solutions were found to be effective in both surgical image and video analysis, where CNN-, LSTM- and CNN-LSTM-based methods were proposed in most surgical video analysis tasks. In this paper, we propose a hybrid spatiotemporal embedding method to enhance spatiotemporal representations using an adaptive fusion layer on top of the LSTM and temporal causal convolutional modules. To learn surgical video representations, we propose exploring the supervised contrastive learning approach to leverage label information in addition to augmented versions. By validating our approach to a video retrieval task on two datasets, Surgical Actions 160 and Cataract-101, we significantly improve on previous results in terms of mean average precision, 30.012 ± 1.778 vs. 22.54 ± 1.557 for Surgical Actions 160 and 81.134 ± 1.28 vs. 33.18 ± 1.311 for Cataract-101. We also validate the proposed method's suitability for surgical phase recognition task using the benchmark Cholec80 surgical dataset, where our approach outperforms (with 90.2% accuracy) the state of the art.

Keywords: laparoscopic video processing; recurrent deep convolutional network; surgical video retrieval; medical multimedia; temporal convolutional network

1. Introduction

In the medical field, interest in minimally invasive surgeries and microscopic surgeries has grown at an enormous rate over the last few decades. Minimally invasive surgery is a surgical technique which requires only small incisions to be made on the patient's body

instead of a large opening, as in case of open surgery. One of the small incisions made is used to insert a small camera called an endoscope, and the surgical instruments needed to perform the operation are inserted through other incisions [1]. Through the endoscope, a high-definition view of the surgical site is displayed on a monitor in the operating room. As a result, the surgeon performs the surgery by looking at a monitor that displays the surgical site inside the patient's body. In the field of medical endoscopy, some of the areas are: esophagogastroduodenoscopy (EGD) (procedures in the gastrointestinal tract), enteroscopy (operations performed on the small intestine), colonoscopy (procedures in the colon), arthroscopy (operations performed on joints), laparoscopy (surgeries performed in the abdomen), etc. The main advantage of the minimally invasive approach over open surgery is that it causes less pain, scarring and patient trauma and it also reduces the risk of infection, hospitalization time and recovery time. On the other hand, microsurgery is a procedure that requires a microscope on small parts of the body. It is used in plastic surgery, cataract surgery, reconstructive surgery involving the skin and muscles, and surgery involving the ear, nose, and throat, etc. In the field of neurosurgery, microsurgery has now become an important process in the treatment of vascular abnormalities as well as cancerous tumors found in the brain.

Nowadays, in the hospital, surgery procedures are often recorded during the operation. In some countries, these videos are also stored in an archive as enforced by law. These recordings can be used in many ways: by young surgeons or medical students to learn the basics of surgery [2], for in-depth analysis of patient disease, for surgical quality assessment [3–5], for analytical self-examination for the surgeon, as an evidence of patients' cases [6], etc. In the field of endoscopic surgery, the endoscope acts as an information source for young trainees who view surgical procedures on a big screen in the surgery room. Additionally, by viewing recorded surgeries later, trainees can further enhance their knowledge. In microscopic surgery, the operating surgeon performs the operation using a microscope, which allows only one trainee to follow the operation with an extra eyepiece and, because of this, there is a hindrance in the teaching and training of young surgeons for this type of surgery. However, microscopic surgery can also be recorded using mounted cameras, and subsequently reviewed in full detail. However, manual searching for the desired video is a tedious and error-prone task. Moreover, both minimally invasive surgery and microscopic surgery typically require specific psychomotor skills that are hard to teach and learn and directly affect the performance of the surgery. These types of surgeries also pose a risk relating to human error due to their nature as high-performance and high-risk undertakings. However, during surgery, these errors and microscopic events may go unnoticed, which prevents the possibility of improvement in future cases. Therefore, it is important to perform surgical error analysis to improve learning and quality control, which will also promote patient safety [7]. With the help of recorded surgery videos, the technical errors during surgery can be carefully examined, and the skill level of the surgeon can also be evaluated [8,9]. One way to use these recordings is to perform a manual search and manual inspection. However, this manual process of searching and inspection is extremely labor-intensive and long-term. Therefore, there is a high requirement for automated content-based surgical video analysis methods for both analysis and searching for desired videos. Consequently, this would allow for investigating other medical research questions in a postoperative manner.

In this work, our main contributions are:

- To support surgical quality assessment, the teaching and training of surgical procedures, and other aforementioned applications, we propose a Content-Based Surgical Video Retrieval (CBSVR) system based on a contrastive learning framework.
- A hybrid temporal embedding approach with adaptive fusion layer is proposed to enhance spatiotemporal features from different modalities.
- We propose a supervised contrastive learning approach to learn surgical video representations which extends the general contrastive loss to consider positive samples from the same label in addition to augmented versions.

- In addition, we design video frame-level self-supervised learning to enhance the visual feature learning when combined with spatiotemporal supervised contrastive learning.
- With extensive experiments, we validate our proposed methodology on two publicly available surgery video datasets along with an ablation study on a surgical phase recognition task.

This paper is organized as follows: Section 2 gives a concise review of related works. Section 3 briefly discusses the essential background information. In Section 4, the proposed framework is elaborated in detail. Experimental settings are discussed in Section 5. The results of the extensive experiments are reported and analyzed in Section 6. Further, the proposed methodology is discussed in Section 7 with experimental outcomes. Finally, Section 8 summarizes the conclusions with future research directions.

2. Related Work

Several works have been published in the last two decades in the field of surgical video analysis [1,10]. Most of the earlier works [11–17] focused on the use of handcrafted features. For instance, the author of [15] used a bag of words approach based on SIFT descriptors for surgical phase recognition in minimally invasive surgery. Reference [16] utilized HOG features for video-based surgical skills assessment. In addition, Allan et al. [17] used SIFT, color-based SIFT and HOG as features for the surgical tool detection task. However, these low-level features are not sufficient to capture subtle details in surgical videos. With the deep learning revolution, recently proposed methods within the realm of medical image analysis appeared to improve feature representation capability through the use of CNN [18–20]. Meanwhile, since a surgical video is in fact a form of sequential data, effectively capturing sequential dynamics in surgical video is important for workflow representation. In this direction, several approaches have also been proposed that are based on statistical models such as hidden Markov models (HMM) [12,21–23], conditional random fields (CRF) [24,25], linear dynamical systems (LDS) [26], dynamic time warping (DTW) [27], etc. Moreover, Cadène et al. [28], Jalal et al. [29], and Twinanda et al. [30] used HMM for modeling temporal information over computed CNN features. However, these statistical methods generally neglect subtle and even significant motions in surgical videos. Recently, some works [31–35] based on RNN have been proposed that work effectively compared to traditional statistical methods. Additionally, works such as [36–38] in the general domain have shown RNN to be effective in video representation.

Diverse attempts have been made in medical image retrieval [39]. However, video retrieval in the medical domain has not yet attracted much attention. Some work has been completed in this field by linking images to corresponding videos. For instance, [40] proposed a video retrieval method based on local and global features (color, texture). The authors of [41] and [42] used feature signatures constructed from the position, color, and texture features to match corresponding videos for querying images. André et al. [43] used the bag of visual words with video-mosaicking technique for endoscopic video retrieval. Furthermore, Beecks et al. [44] proposed gradient-based signatures for image and video linking. However, these static frame-based methods ignore the crucial temporal dynamics present in the video, affecting system performance. Therefore, it is important to incorporate motion information present in the video [45,46]. For instance, Syeda-Mahmood et al. [47], Quillec et al. [48] and Quillec et al. [49] used optical flow for motion information. Droueche et al. [50] proposed motion trajectory-based method for content-based medical video retrieval. They used motion vectors from the MPEG-4 video stream to build trajectories and used the extension of DTW for video matching. Muenzer et al. [51] investigated static frame-based features and dynamic-based features for laparoscopic video retrieval, and also gave directions to fuse these features for better representation. Kletz et al. [52] proposed flow-based and track-based descriptors for similarity searches that extend static frame-based feature signatures to dynamically rich content. Moreover, Amanat et al. [53] proposed a meniscal surgical video retrieval system based on the use of key points, statistical, PCA-SIFT and PCA-GMM methods, and found that PCA-SIFT-based performance

is better than others. In this method, a master frame is first extracted by a shot boundary detector and then its SIFT description is computed followed by PCA for dimensionality reduction. The author used Euclidean distance to match the input video to target video. Furthermore, Schoeffmann et al. [54] proposed a video content descriptor for laparoscopic video retrieval named motion intensity and direction descriptor (MIDD). In their proposed method, they first estimated motion vectors using optical flow and then by quantizing these motion vectors into different bins, they computed a motion direction histogram and used it as a descriptor to represent the surgical sequence. In contrast to the above methods, our approach explicitly exploits both spatial and temporal information present in the surgical video without using optical flow under the unified single model. Additionally, we explore the hybrid embedding learning approach to enhance spatiotemporal representations.

3. Background

Deeper CNNs typically take longer to converge during backpropagation, which is due to the rapid shrinking of the gradients toward input while backpropagating the loss, and thus require a large number of iterations to update the parameters. He et al. [55] proposed ResNet that overcomes the issue of training deeper CNNs by the use of a residual function. In our proposed method, we follow the ResNet-18, which acts as an initial visual descriptor. In general, it takes a 224×224 RGB image/frame and transforms through a number of residual blocks to generate 512-dimension activation at the end by means of an average pooling layer.

On the other hand, RNN is a special type of neural network that processes sequential data. In general, as shown in Figure 1a, RNN processes the sequential data by looking at the input as well as the output of the hidden state of the previous time step to synthesize dynamics in data. However, for long sequences, the gradients can grow and decay exponentially during training, which results in its limitation to learn long-length continuous dynamics. To address this problem, LSTM is developed in [56] and emerged as successful architecture in tasks related to natural language processing [57]. LSTM (see Figure 1b) uses memory cells in long-range learning that tell the network what to remember and when to forget. We use the LSTM module for learning temporal feature representation. The major advantage of the differentiable recurrent models such as LSTM is that they can map variable-length input such as video to single-label output (e.g., genre category) directly and can also model complex temporal dynamics as demonstrated in [36].

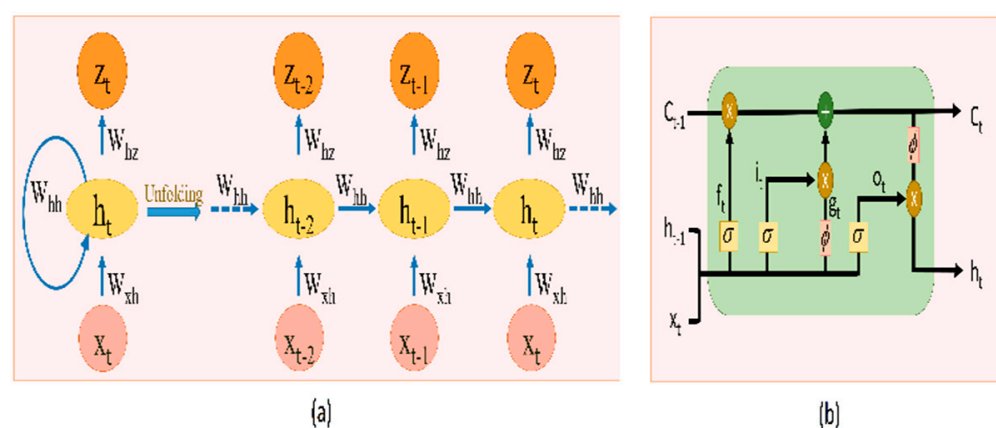


Figure 1. (a) Unfolded version of basic RNN; (b) LSTM module.

Temporal relationships can also be synthesized with temporal convolutions [58]. For instance, [59] exploited residual 1D temporal convolutions and proposed a multi-stage temporal convolutional network for surgical phase recognition. Additionally, Ramesh et al. [60] proposed a multi-stage temporal convolutional network for the recognition of phases in gastric bypass procedures. These methods are based on causal convolutions, i.e., the prediction at current time step does not involve future information but only past

information. This is achieved by convolving with applying padding at the single end, i.e., at the past side. In Figure 2, the visualization of the causal convolution is shown. We used the causal convolutions in our temporal convolution layer (with kernel size 1×3) to capture temporal information in hybrid mode.

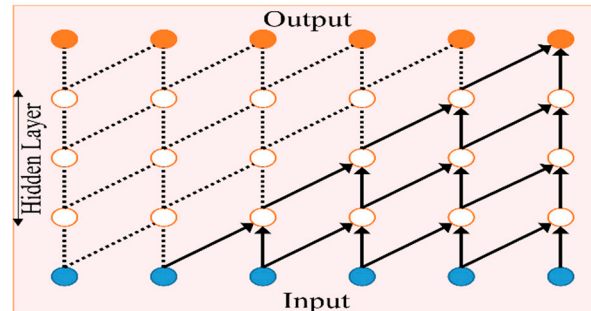


Figure 2. Causal convolution.

Recently, contrastive learning [61] has emerged as an effective method for learning robust representations in the image field. The general concept is to make an anchor-positive pair closer in the embedding space, and to separate the anchor from the negative samples. A deep network such as ResNet50 is used as an encoder to extract high-level features (of dimension 2048), which are then projected to embedding space with a projection head (with lower dimension, e.g., 256). As in [61], when no labels are used, positive pairs are constructed with data augmentations, and negatives for the anchor are sampled randomly from the mini-batch.

4. Proposed Methodology

The proposed CB-SVR system including training and retrieval process is illustrated in Figure 3. The system has three core elements: video representation learning, feature extraction and the retrieval process. Firstly, video representation learning is performed by optimizing the weights in the proposed hybrid feature embedding learning model. Second, video-level features are extracted by deploying the trained model to index the database. Finally, similar surgery videos are retrieved for a given query surgical video by matching the query's video descriptor to the stored database's index of features via similarity measurement. All of the elements of the proposed framework are detailed in the following subsections.

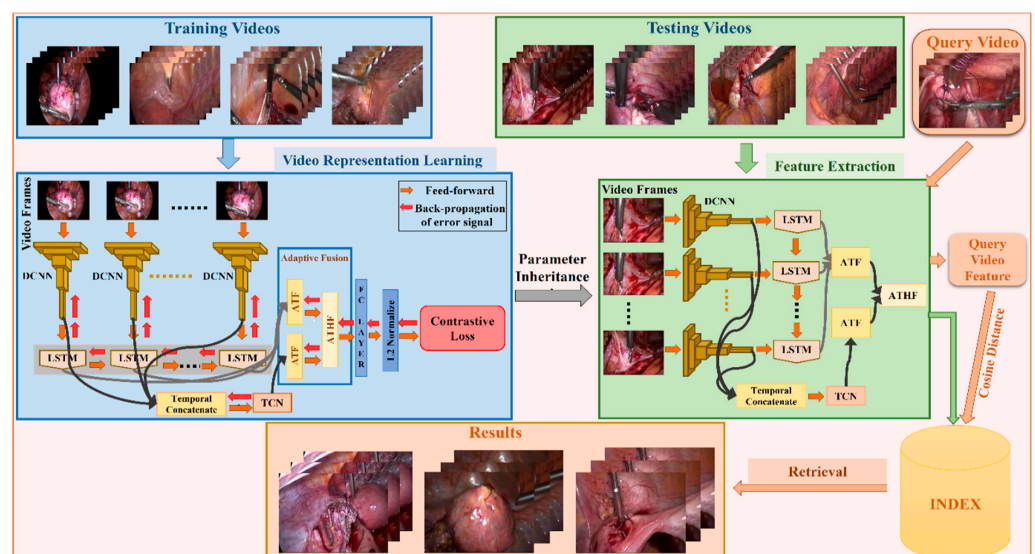


Figure 3. Overview of the proposed CB-SVR system.

4.1. Video Representation Learning

Given a surgical video sequence $V^{(j)} = [V_1^{(j)}, V_2^{(j)}, \dots, V_t^{(j)}, \dots, V_n^{(j)}] \in R^{n \times r \times c}$ (where $V^{(j)}$ indicates the j th surgical video in the database, n refers to the number of frames in each video, and $r \times c$ is a spatial resolution of each frame), the goal of the video representation learning method is to learn a compact descriptor or code for each surgical video. In order to learn video representation directly from the surgical video itself, we adapt the deep learning approach to learn long-range dynamics in a surgical video through the proposed hybrid spatiotemporal embedding learning method.

4.1.1. Hybrid Spatiotemporal Contrastive Embedding Learning

This work proposes a hybrid spatiotemporal embedding learning (RDCN–TCN–CL) method within a supervised contrastive learning framework (see Figure 4) to learn surgical video representations. Inspired by [61], we propose exploring the supervised contrastive learning method in the context of surgical video retrieval, where we leverage the label information to construct positive pairs in addition to augmentations. We design the hybrid model RDCN–TCN–CL by combining CNN as spatial encoder and LSTM and temporal causal convolution modules for temporal information analysis. We design an adaptive temporal fusion layer which aims to pool the sequence of spatiotemporal features with learnable weights. The pooled features are then fused together via an adaptive hybrid fusion layer, which aims to fuse spatiotemporal features of different temporal modules by applying appropriate weights. The entire model is trained under contrast learning which includes video frame embedding learning and spatiotemporal embedding learning.

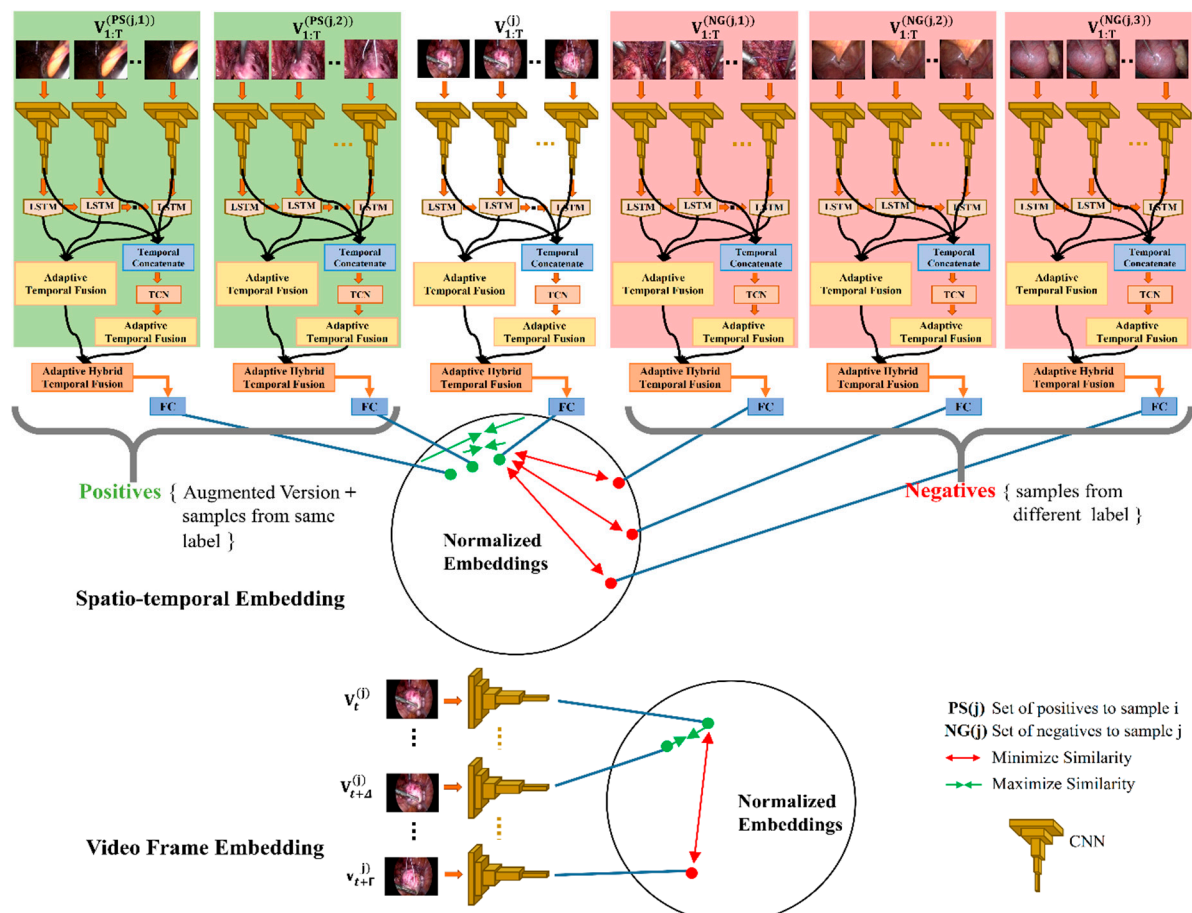


Figure 4. Proposed hybrid spatiotemporal contrastive embedding learning for surgical video.

The RDCN–TCN–CL works by feeding T continuous frames of input surgical video $V_{1:T}^{(j)}$ into T shared deep CNNs (ResNet-18 in our case) to encode spatial information within the pixels into rich features as a sequence of frame-level features $S_{1:T}^{(j)}$ which are then input to the recurrent sequence learning module (LSTM) to learn temporal dependencies in the surgical video. Specifically, a sequence of visual features $S_{1:T}^{(j)}$ is generated by passing each time step $t \leq T$ of video sequence $V_t^{(j)}$ through a feature transformation f_{DCNN} with learnable parameter U as $S_t^{(j)} = f_{DCNN}(V_t^{(j)}, U)$. The transformation f_{DCNN} corresponds to the activations of the last layer (average pooling layer in our case) before the classification layer in the deep CNN. The sequence $S_{1:T}^{(j)}$ is then input sequentially into a single-layer LSTM with 512 hidden units, where the cell state c_t and hidden state h_t of LSTM for time step t are updated as:

$$i_t = \sigma(Q_{si} S_t + Q_{hi} h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(Q_{sf} S_t + Q_{hf} h_{t-1} + b_f) \quad (2)$$

$$g_t = \phi(Q_{sg} S_t + Q_{hg} h_{t-1} + b_g) \quad (3)$$

$$o_t = \sigma(Q_{so} S_t + Q_{ho} h_{t-1} + b_o) \quad (4)$$

$$c_t = (f_t \odot c_{t-1} + i_t \odot g_t) \quad (5)$$

$$h_t = o_t \odot \phi(c_t) \quad (6)$$

where Q and b are the LSTM's parameters; i , g , o are the input gate, input modulation gate and output gate, respectively; f acts as the forget gate where c acts as the memory cell; σ and ϕ are the sigmoid and hyperbolic tangent function, respectively; and \odot refers to element-wise multiplication.

To further enrich the feature descriptor of the surgical sequence $S_{1:T}^{(j)}$, the temporal causal convolutional layer is included in the proposed framework, consisting of 5 layers, each consisting of 64 kernels with a kernel size of 1×3 . The output of the temporal convolutional layer is the same size as the input $S_{1:T}^{(j)}$ and denoted as $G_{1:T}^{(j)}$.

Adaptive Fusion

For effective analysis of the surgical video content, the use of all hidden states of the LSTM module as well as all time steps of the causal convolution module are critical. In general, all segments of the video have different effects, and the simple temporal mean pooling operation can ignore the important weights of different segments. To tackle this, we design the adaptive temporal fusion (ATF) layer to learn these weights.

Let $h_{1:T}$ be the hidden states of the LSTM module and $G_{1:T}^{(j)}$ be the output of the temporal causal convolutional layer. The output of the ATF layer for both modules is computed as:

$$ATF_{LSTM}^{(j)} = \sum_{t=1:T} (h_t^{(j)} \odot a1_t) / T \quad (7)$$

$$ATF_{TCN}^{(j)} = \sum_{t=1:T} (G_t^{(j)} \odot a1_t) / T \quad (8)$$

where $a1_t$ is the adaptive weight and \odot denotes element-wise multiplication.

$ATF_{LSTM}^{(j)}$ and $ATF_{TCN}^{(j)}$ represent clip-level descriptors from two different temporal modules. For effective fusion of these two features, we design the adaptive hybrid temporal fusion (AHTF) layer. It consists of 1 fully connected layer. The output of AHTF layer is computed as:

$$AHTF^{(j)} = \left[\left((W_{AF} ATF_{LSTM}^{(j)} + b) \odot a2 \right) \parallel \left((W_{AF} ATF_{LSTM}^{(j)} + b) \odot a3 \right) \right] \quad (9)$$

where WAF is the shared weight matrix, b is the bias term, and $a2$ and $a3$ are adaptive weights for the $LSTM$ and the temporal causal convolutional modality.

Contrastive Embedding Learning

The feature $AHTF^{(j)}$ is passed through the fully connected layer $z = W_{HF}AHTF^{(j)} + b$ (where $W_{HF} \in R^{|K| \times d}$ and $b \in R^{|K|}$ are learnable parameters) followed by a L2 normalized layer.

Now, let z_j be the L2 normalized feature vector corresponding to surgical video $V_{1:T}^{(j)}$, and consider a positive pair (z_j, z_{pos}) from the same class (and augmented versions) and negatives (z_j, z_{neg}) from different classes within the training minibatch, the probability of z_j being recognized as the category of its positive sample z_{pos} as:

$$P(y_j = y_{pos} | z_j) = \frac{\exp(z_{pos}^T z_j / \tau)}{\exp(z_{pos}^T z_j / \tau) + \sum_{neg \in NG(j)} \exp(z_{neg}^T z_j / \tau)} \quad (10)$$

where $NG(j)$ is a set of negatives to the j th sample and τ is temperature [61].

Let N refer to the training set of labeled surgical video sequences, i.e., $(V_{1:T}^{(j)}, y^{(j)}) \in N$, where y is the label. Let PS and NG be sets of positive and negative samples. The goal is to maximize the probability (10). This can be accomplished by optimizing the parameters (U, Q) of the network's visual and sequential components by minimizing the negative log likelihood. The contrastive loss associated with (11) can be given as

$$L_{VCL} = \frac{1}{|N|} \sum_{j=1}^N \frac{1}{|PS(j)|} \sum_{pos \in PS(j)} (-\log P(y_j = y_{pos} | z_j)) \quad (11)$$

which can be rewritten as:

$$L_{VCL} = \frac{1}{|N|} \sum_{j=1}^N \frac{1}{|PS(j)|} \sum_{pos \in PS(j)} \left(-\log \left(\exp(z_{pos}^T z_j / \tau) / \exp(z_{pos}^T z_j / \tau) + \sum_{neg \in NG(j)} \exp(z_{neg}^T z_j / \tau) \right) \right) \quad (12)$$

By minimizing Equation (12), we end up in maximizing the cosine similarity between z_j and z_{pos} and minimizing the cosine similarity between z_j and z_{neg} , where $y_j = y_{pos}$ and $y_j \neq y_{neg}$.

In addition to video-level contrastive loss (12), we also apply contrastive learning at the frame level so that it can better learn the visual patterns essential for video representation. We design the self-supervised learning task with the assumption that adjacent frames share similar patterns with respect to distant frames. More formally, the required constraints for embedding learning are:

$$f_{DCNN_t}(V_t^{(j)}, U) \approx f_{DCNN_{t+\Delta}}(V_{t+\Delta}^{(j)}, U)$$

for small Δ , and

$$Dist(f_{DCNN_t}(V_t^{(j)}, U), f_{DCNN_{t+\Delta}}(V_{t+\Delta}^{(j)}, U)) < Dist(f_{DCNN_t}(V_t^{(j)}, U), f_{DCNN_{t+\Gamma}}(V_{t+\Gamma}^{(j)}, U))$$

for $\Gamma > \Delta$, where $Dist$ is the distance function (we used the L2 norm).

The contrastive loss for video frame embedding can be defined as:

$$L_{FCL} = Dist(f_{DCNN_t}(V_t^{(j)}, U), f_{DCNN_{t+\Delta}}(V_{t+\Delta}^{(j)}, U)) + \max\{0, m_c - Dist(f_{DCNN_t}(V_t^{(j)}, U), f_{DCNN_{t+\Gamma}}(V_{t+\Gamma}^{(j)}, U))\} \quad (13)$$

The loss L_{FCL} will make $f_{DCNN_t}(V_t^{(j)}, U)$ and $f_{DCNN_{t+\Delta}}(V_{t+\Delta}^{(j)}, U)$ closer, while $f_{DCNN_t}(V_t^{(j)}, U)$ and $f_{DCNN_{t+\Gamma}}(V_{t+\Gamma}^{(j)}, U)$ are enforced to be separated by the margin m_c .

The total loss is given as:

$$L_{CL} = L_{VCL} + \lambda L_{FCL} \quad (14)$$

We jointly optimize the entire parameters of RDCN–TCN–CL, with backpropagation used to compute the gradient of the objective L_{CL} with respect to all parameters (U, Q) over minibatches $N' \subset N$ sampled from the training dataset N .

4.1.2. Training Details

Since the parameter scale of the spatial component of RDCN–TCN–CL is much larger than that of its temporal component, it can quickly overfit on smaller training datasets. Additionally, the effectiveness of its temporal component depends on how its spatial component extracts the features that are relevant to its temporal component. Therefore, to effectively train the RDCN–TCN–CL, we initialize the weights of its spatial component with weights trained on the ImageNet [62] dataset by leveraging the power of transfer learning. Further, we initialize the recurrent weights of its LSTM module with orthogonal initialization [63] and the TCN module as in [55]. The weights of adaptive fusion layer are initialized to 1. We train the RDCN–TCN–CL network using backpropagation with Adam optimization [64] with β_1 set to 0.9, β_2 set to 0.999 and $\epsilon = 10 \times 10^{-8}$, and weight decay of 10×10^{-4} . The neuron size of embedding is set to 256. The learning rate is set to 1×10^{-4} . The parameter Δ is set to choose randomly from $\{t = 0:\text{floor}(T/4)\}$ time step and Γ from $\{t = T-\text{floor}(T/4):T\}$. Additionally, m_c is set to 2. Furthermore, in order to prevent overfitting, we also adopt undermentioned data augmentation techniques and an early stopping strategy. We stop training after 100 epochs, as the loss does not seem to decrease further.

During training, for each training batch, we apply three types of data augmentation, namely cropping, rotation and horizontal flipping, to artificially enlarge the database. First, a video clip of 24 (i.e., $T = 24$ without temporal downsampling) continuous frames is randomly sampled from each video, after which it is converted to a 250×280 spatial resolution. Then, the center crop of 240×260 is sampled from its randomly rotated variant within a range of $[-5, 5]$. After that, the 224×224 crop is randomly sampled from it and fed to the network. We also apply horizontal flipping with 50% probability before inputting it into network. Our framework is implemented on MATLAB 2019b with NVIDIA Tesla K40c GPU.

4.2. Feature Extraction and Query Matching

Once the model is trained, its responses can be used as a feature representation for the surgical video. The features from the LSTM component and TCN component in the trained RDCN–TCN–CL model are extracted and fused via the adaptive fusion layer to represent the surgical sequence.

To facilitate the retrieval process, the model's responses as features are extracted from each surgical video in the database using (9) and indexed in the database. Now, for a given query surgical video q , which is represented as a feature vector obtained after feature extraction, CB-SVR aims to select the X best videos from the database that resembles this query video q . This is usually done by computing the similarity distance between the query video and the videos in the database D . In this paper we utilize the cosine distance as similarity measurement which is given as:

$$D_{\cos}(q, D) = \frac{\sqrt{\sum_{r=1}^d (F_{q_r})^2} \sqrt{\sum_{r=1}^d (F_{D_{jr}})^2} - \sum_{r=1}^d F_{q_r} F_{D_{jr}}}{\sqrt{\sum_{r=1}^d (F_{q_r})^2} \sqrt{\sum_{r=1}^d (F_{D_{jr}})^2}} \quad (15)$$

where F_{q_r} is the r th feature vector of query video q and $F_{D_{jr}}$ is the r th feature of the j th video in the database.

5. Dataset

We chose two publicly available datasets to evaluate the performance of our method: Surgical Actions 160 [54] and Cataract-101 [65]. The Surgical Actions 160 dataset consists of 160 short video clips representing surgical actions in gynecologic laparoscopy. These video clips are categorized into 16 distinct classes of surgical actions as depicted in Figure 5, and the database has exactly 10 example clips for each action class. Each video clip is encoded with H.264/AVC and is of approximately 5-second duration with a spatial resolution of 320×240 . In total, this dataset amounts to 19,181 frames. For our experiments, we randomly selected ten splits, such that each split divided the dataset into two sets (training and testing), and each set consists of 80 videos (5 videos per class in each set).

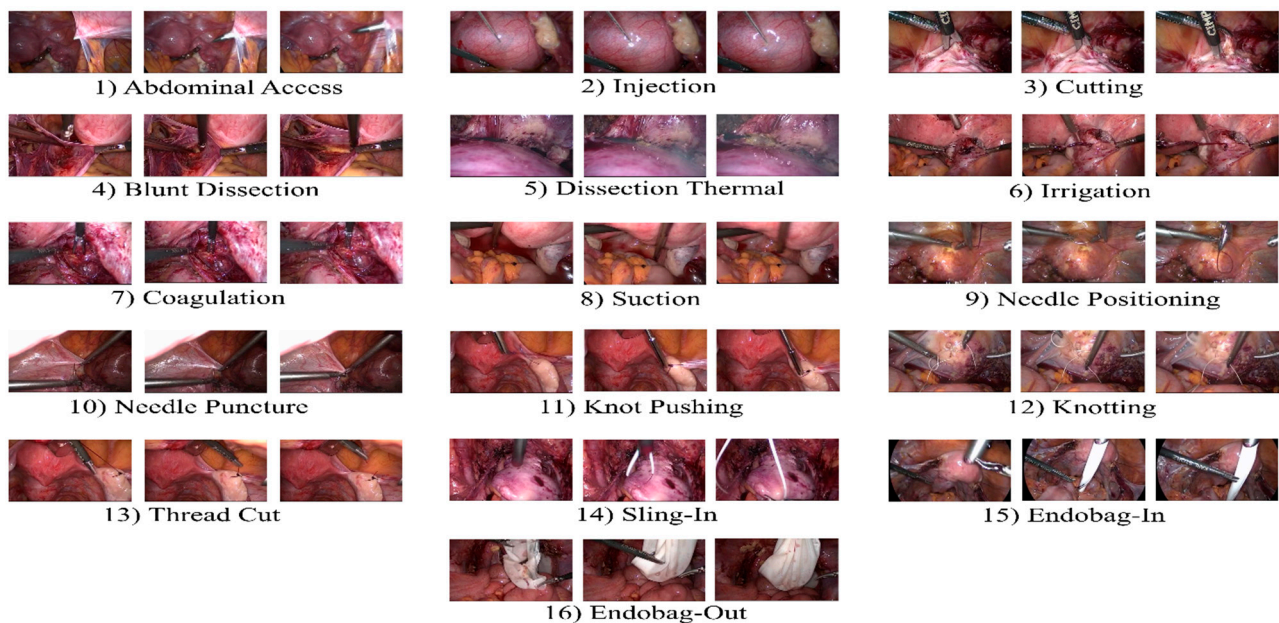


Figure 5. Overview of 16 different surgical actions in gynecologic laparoscopy.

On the other hand, the Cataract-101 dataset consists of 101 recorded cataract surgeries performed by four surgeons [65] and annotated with ten surgical phases (as shown in Figure 6) by the senior ophthalmic surgeon. Each video is encoded with H.264/AVC. In total, this dataset amounts to 1.26 million frames with a total duration of 14 h and a spatial resolution of 720×540 . For our experiments, we prepared this dataset for the experiment as follows: first, we extracted clips from each video according to the phase annotation given with the dataset, resulting in 1265 total clips of 10 phases. Then, we randomly chose 5 splits such that for each split, the training set and testing set consisted of 634 and 631 video clips, respectively, as depicted in Table 1.

Training is performed using a training set and retrieval is performed on the testing set with the leave one out rule (i.e., retrieval set does not include query video). As visible in both Figures 5 and 6, the visual appearance of both datasets is very similar, which makes retrieval of surgical videos an extremely challenging research area.

We adopted mean average precision (mAP) following [54] to evaluate the effectiveness of the proposed approach, which is computed using (16).

$$\text{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \left(\frac{1}{G_q} \sum_{k=1}^{G_q} P_q(k) \beta_q(k) \right) \quad (16)$$

where G_q is the ground truth on query q . $P_q(k)$ is the precision of the top k retrieved videos, Q is the set of all queries and $\beta_q(k) = 1$ if the item at rank k is relevant, and otherwise 0.

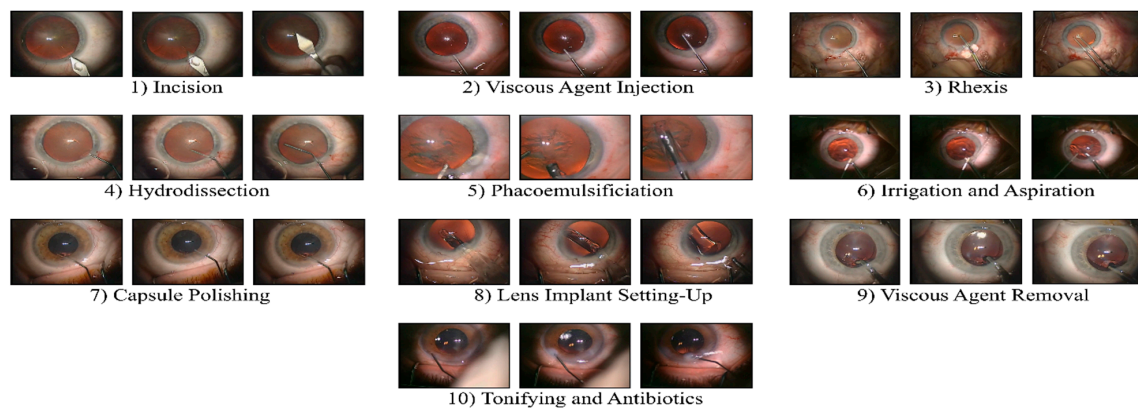


Figure 6. Overview of 10 different surgical phases in cataract surgery.

Table 1. Dataset (Cataract-101) preparation.

Surgery Phase/Label	#Videos		
		Training	Testing
Incision	104	52	52
Viscous agent injection	232	116	116
Rhexis	104	52	52
Hydrodissection	101	51	50
Phacoemulsification	104	52	52
Irrigation and aspiration	120	60	60
Capsule polishing	110	55	55
Lens implant setting-up	105	53	52
Viscous agent removal	106	53	53
Tonifying and antibiotics	179	90	89
TOTAL	1265	634	631

6. Experimental Results and Analysis

6.1. Analysis of Training with Different Temporal Length Video Sequences

We first investigate the influence of learned RDCN–TCN–CL features on the retrieval performance by selecting different temporal lengths ($T = \{8, 16, 24, 32\}$) of video sequence inputs for training the RDCN–TCN–CL. As shown in Table 2, the mAP (averaged over aforementioned splits) gradually increases with an increase in the temporal length of the input sequence. This indicates that a longer input sequence for training gives a boost in system performance. However, a input sequence length of 32 does not have a further impact on the retrieval accuracy but costs more during training. Hence, we choose the temporal length of 24 for a video sequence for training the RDCN–TCN–CL network under the CB-SVR system. Note that we do not downsample the video and use the original sampling. Training with downsampled sequence (to increase the length of video) can be further investigated.

Table 2. mAP % (mean \pm std.) when training under different temporal lengths of video sequence on two datasets under the CB-SVR system.

Dataset	Temporal Factor			
	8	16	24	32
Surgical Actions 160	27.11 \pm 1.21	27.89 \pm 1.12	30.012 \pm 1.778	30.09 \pm 1.14
Cataract-101	78.43 \pm 0.673	79.25 \pm 0.556	81.13 \pm 1.28	81.11 \pm 1.08

6.2. Analysis of Different Temporal Pooling Strategies

Extracting discriminative video features is vital for our task. In this regard, we evaluate the impact of two feature pooling methods (mean pool, max pool) on retrieval performance. For this assessment, video features are computed by using with particular pooling method in place of adaptive fusion. The retrieval performance (mAP averaged over aforementioned splits) based on these methods is depicted in Table 3. The result shows that adaptive fusion performs superiorly over the other two pooling methods with 30.012% mAP in Surgical Actions 160 and 81.13% mAP in Cataract-101.

Table 3. mAP (averaged over aforementioned splits) of different feature pooling strategies on two datasets under the CB-SVR system.

Dataset	Max Pool	Mean Pool	Adaptive Fusion
Surgical Actions 160	27.64 ± 1.17	28.63 ± 1.76	30.012 ± 1.778
Cataract-101	76.71 ± 1.21	79.72 ± 0.28	81.13 ± 1.28

6.3. Effectiveness of Combined Visual and Sequential learning

In order to demonstrate the effectiveness of RDCN–TCN–CL as a combined visual and sequential features learnable network, we evaluate its performance with a series of experiments by exploiting ResNet-18 into its potential counterparts. These are: (1) ResNet18_S: ResNet-18 trained from scratch; (2) ResNet18_{FT}: ResNet-18 fine-tuned from an imagenet pre-trained model; (3) ResNet18_{FRZ} + LSTM: trained on frame-level features computed from a frozen imagenet pre-trained model; (4) ResNet18_{FRZ} + LSTM + TCN: trained with hybrid LSTM–TCN on frame-level features computed from a frozen imagenet pre-trained model; and (5) RDCN – TCN – CL_S: ResNet-18, LSTM and TCN jointly trained end-to-end from scratch. We use the same training parameters for ResNet18_S, ResNet18_{FT}, ResNet18_{FRZ} + LSTM, ResNet18_{FRZ} + LSTM + TCN and RDCN – TCN – CL_S as stated in Section 4.1.2 for RDCN–TCN–CL. We initialize the weights of ResNet18_S and the spatial component of RDCN – TCN – CL_S as in [55]. Moreover, we train ResNet18_S and ResNet18_{FT} in its purest form, i.e., we do not use dropout, but we follow early stopping and all spatial data augmentation techniques (with temporal length of 1) to avoid overfitting as stated in Section 4.1.2. The video features are extracted with spatial frame-level features (extracted from respected average pooling layer) for ResNet18_S and ResNet18_{FT}. Likewise, for ResNet18_{FRZ} + LSTM, ResNet18_{FRZ} + LSTM + TCN and RDCN – TCN – CL_S, it follows the same procedure to extract video features as in RDCN–TCN–CL. All counterparts with RDCN–TCN–CL are performed on the aforementioned training and testing splits, and the results are depicted in Tables 4 and 5 for the Surgical Actions 160 and Cataract-101 databases, respectively.

Table 4. Comparison between RDCN–TCN–CL and other network architectures in terms of retrieval performance (mAP \pm std.) on Surgical Actions 160 under the CB-SVR system.

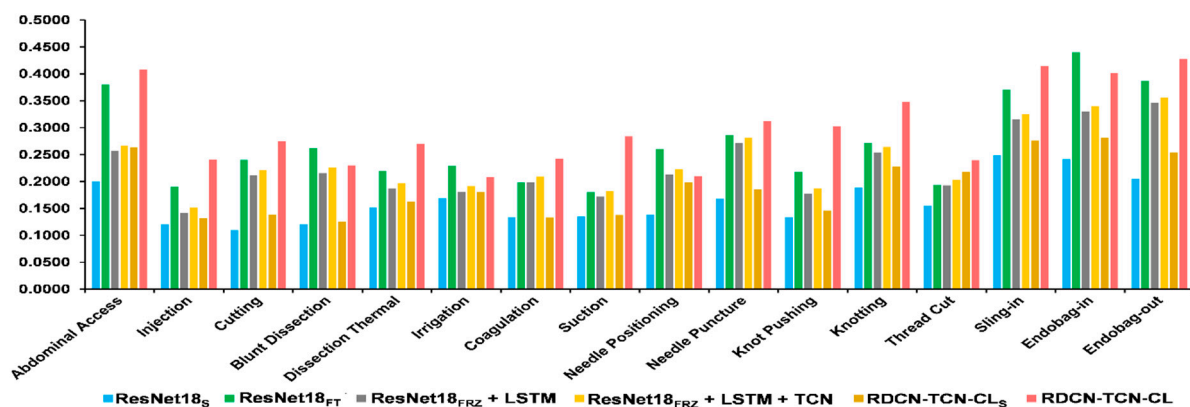
Split	ResNet18 _S	ResNet18 _{FT}	ResNet18 _{FRZ} + LSTM	ResNet18 _{FRZ} + LSTM + TCN	RDCN – TCN – CL _S	RDCN–TCN–CL
1	14.71	25.59	18.52	18.81	14.78	28.62
2	19.35	30.47	20.87	21.42	18.14	32.34
3	14.97	25.14	19.14	19.25	14.21	28.67
4	14.36	26.37	18.12	18.42	13.98	27.78
5	17.42	28.16	22.42	22.51	17.88	30.71
6	15.65	26.17	21.13	22.24	15.65	29.42
7	16.75	29.18	22.21	23.11	17.23	31.58
8	14.41	26.09	20.17	21.12	14.34	27.48
9	18.27	26.74	21.36	22.46	19.14	32.67
10	17.24	26.11	21.54	21.82	17.21	30.85
Mean \pm std.	16.31 ± 1.660	27.002 ± 1.623	20.548 ± 1.431	21.12 ± 1.601	16.256 ± 1.787	30.012 ± 1.778

Table 5. Comparison between RDCN–TCN–CL and other network architectures in terms of retrieval performance (mAP \pm std.) on Cataract-101 under the CB-SVR system.

Split	ResNet18 _S	ResNet18 _{FT}	ResNet18 _{FRZ} + LSTM	ResNet18 _{FRZ} + LSTM + TCN	RDCN – TCN – CL _S	RDCN–TCN–CL
1	46.48	73.75	53.89	54.32	55.26	81.35
2	46.25	74.66	54.25	55.29	57.45	79.25
3	42.56	73.02	53.83	54.96	50.96	82.73
4	48.68	74.67	54.13	56.01	52.90	80.14
5	51.35	75.81	54.37	56.15	51.96	82.2
Mean \pm std.	47.06 \pm 2.91	74.38 \pm 0.9438	54.27 \pm 0.327	55.35 \pm 0.677	53.70 \pm 2.352	81.134 \pm 1.28

In Tables 4 and 5, we can see that RDCN–TCN–CL achieves much better performance than the other four methods. More analytically, in the context of retrieval performance (mAP averaged over aforementioned splits), we can observe: (1) the importance of pre-trained weights (in initializing the spatial component DCNN) when comparing ResNet18_S with ResNet18_{FT}, and RDCN – TCN – CL_S with RDCN–TCN–CL; (2) the importance of using temporal dynamics when comparing ResNet18_S and RDCN – TCN – CL_S (although there is a little difference in their performance for the Surgical Actions 160 dataset, which may be due to there being fewer training data than Cataract-101); (3) the effectiveness of end-to-end training when comparing ResNet18_{FRZ} + LSTM and RDCN–TCN–CL; and (4) despite using both spatial and temporal dynamics, ResNet18_{FRZ} + LSTM performs worse than ResNet18_{FT}. The underlying reason for this is that ResNet18_{FT} is a domain-specific version adapted to the surgical visual content, whereas ResNet18_{FRZ} + LSTM relies on visual knowledge of the general domain and, therefore, is unable to synthesize true temporal dynamics in the surgical video.

Further, Figures 7 and 8 depict the class-wise performance, where RDCN–TCN–CL performs much better than other methods in all classes of Cataract-101, but in Surgical Actions 160, we can see its performance degradation in some classes (e.g., Blunt Dissection, Irrigation, Needle Positioning and Endobag-in). However, for such classes, ResNet18_{FT} performs better. The reason for this seems to be that these classes have some unique visual patterns in the form of instruments and structural objects that are effectively captured with static ResNet18_{FT} descriptors, while for other classes which have strong subtle motions, RDCN–TCN–CL descriptors are able to capture them.

**Figure 7.** Class-wise comparison of Surgical Actions 160 between RDCN–TCN–CL and other network architectures, in the terms of retrieval performance (mAP averaged over 10 splits).

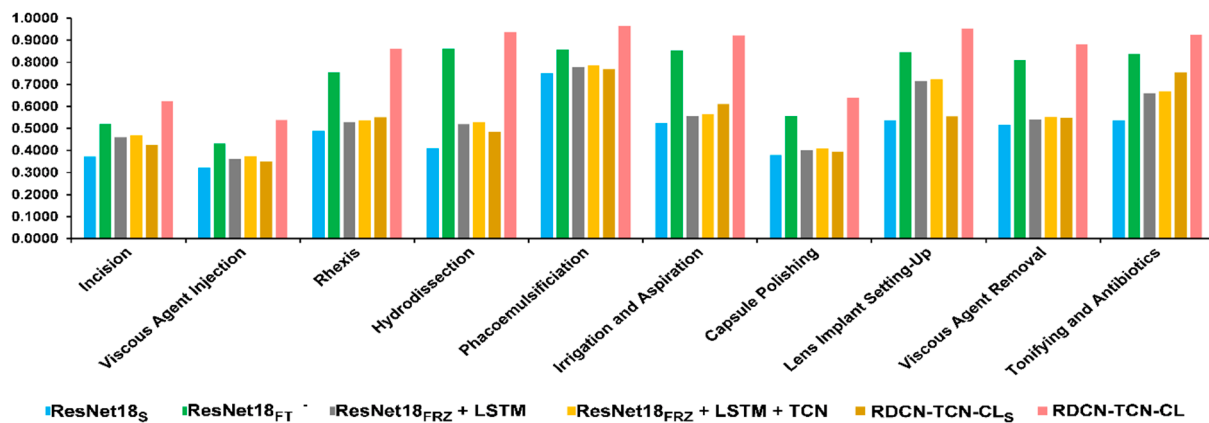


Figure 8. Class-wise comparison of Cataract-101 between RDCN-TCN-CL and other network architectures, in the terms of retrieval performance (mAP averaged over 5 splits).

6.4. Comparison with State of the Art

In Table 6, to demonstrate the superiority of the proposed methodology, we compare our approach with previous methods on both datasets. We also compute the CNN responses such as ResNet18 responses from the average pooling layer as a feature descriptor (denoted as CNNR). Similarly, CNNA represents activations of the fc7 layer of the AlexNet [66], and CNNG represents Googlenet [67] responses from the last pooling layer. For the CNN features, we perform meanpool in the temporal dimension.

Table 6. Comparison with the state of the art on Surgical Actions 160.

Feature Descriptor	Dim.	mAP (Mean \pm std.) Surgical Actions 160	Cataract-101
CNNA [66]	4096	20.011 \pm 1.531	20.44 \pm 1.253
CNNG [67]	1024	21.82 \pm 1.064	22.16 \pm 1.145
CNNR [55]	512	22.67 \pm 1.192	23.18 \pm 1.212
FS [41]	630	19.46 \pm 1.631	20.11 \pm 1.264
DFS [54]	810	20.75 \pm 1.253	25.18 \pm 1.152
MIDD [54]	25	22.54 \pm 1.557	33.18 \pm 1.311
RDCN-TCN-CE	512	28.533 \pm 1.142	75.27 \pm 1.35
RDCN-TCN-Triplet	512	28.741 \pm 1.334	77.36 \pm 1.18
RDCN-TCN-CL	512	30.012 \pm 1.778	81.134 \pm 1.28

In Table 7, we can observe the impact of network depth on the retrieval performance using the results obtained by CNNA (20.011 \pm 1.53)%, CNNG (21.82 \pm 1.06)%, and CNNR (22.67 \pm 1.19)% for the Surgical Actions dataset and CNNA (20.44 \pm 1.253)%, CNNG (22.16 \pm 1.145)%, and CNNR (23.18 \pm 1.212)% for the Cataract-101 dataset. Furthermore, the handcrafted feature-based method [41] performs poorly compared to deep features (e.g., 19.46 \pm 1.631 vs. 22.67 \pm 1.192) for surgical 160). With RDCN-TCN-CL descriptor, we are able to outperform MIDD [54] with an mAP of 30.012 \pm 1.778 in Surgical Actions 160 and with a significant boost in mAP (81.13 \pm 1.28 vs. 33.18 \pm 1.311) for Cataract-101. Additionally, RDCN-TCN-CL was found to be effective compared to RDCN-TCN-Triplet and RDCN-TCN-CE, where RDCN-TCN-Triplet is trained under triplet loss and RDCN-TCN-CE is trained under cross entropy loss.

Table 7. Surgical workflow recognition performance on Cholec80.

Methods	Accuracy	Precision	Recall
PhaseNet [30]	78.8 ± 4.7	71.3 ± 15.6	76.6 ± 16.6
EndoNet * [30]	81.7 ± 4.2	73.70 ± 16.1	79.60 ± 7.9
(EndoNet + LSTM) * [68]	88.6 ± 9.6	84.4 ± 7.9	84.7 ± 7.9
SV-RCNet [32]	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5
MTRCNet * [33]	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9
TeCNO [59]	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7
NL-RCNet [34]	85.73 ± 6.96	82.94 ± 6.20	85.04 ± 5.15
RDCN–TCN–CE (R18)	82.13 ± 6.7	75.14 ± 9.1	79.21 ± 7.1
RDCN–TCN–Triplet (R18)	83.04 ± 7.1	77.65 ± 10.5	80.43 ± 7.4
RDCN–TCN–CL (R18)	83.86 ± 6.67	78.89 ± 7.3	81.15 ± 7.2
RDCN–TCN–CL (R50)	90.2 ± 6.93	87.52 ± 6.88	85.65 ± 6.91

* means methods of multi-task learning with extra tool labels needed.

We further evaluate run time performance of RDCN–TCN–CL on Xeon E5 CPU. Compared to MIDD, which can process 108 frames per second for feature extraction, RDCN–TCN–CL can process 40 frames per second. Although RDCN–TCN–CL is slower than MIDD, it achieves higher performance than MIDD, and is a promising future approach.

6.5. Ablation Study in Context of Surgical Phase Recognition

Next, we also conduct an ablation study to analyze the effectiveness of RDCN–TCN–CL in surgical phase recognition. In this regard, we conduct the experiment on a large dataset named cholec80 dataset [68] and compared the performance with the state of the art. Cholec80 consists of eighty recorded videos of cholecystectomy surgeries captured at 25 fps. All videos are annotated with seven phases and seven tools. Following [32,33,68], we used 40 videos as a training set and the rest as a test set. Since additional annotations such as tool presence labels are generally not available and single-task approaches are more practical in real-world applications, we only employ the phase labels. Further, we downsample the training videos from 25 fps to 1 fps. The video frames are first resized to 250 × 250 and follow the same data augmentation as mentioned in Section 4.1.2.

The loss for the surgical phase recognition task with contrastive loss can be defined as:

$$L_P = \frac{1}{|N|} \sum_{j=1}^N \sum_{t=1}^T \left(-\log(\hat{y}_t^{(j)}) \right) + \frac{1}{|N|} \sum_{j=1}^N \sum_{t=1}^T \left(-\log(\bar{y}_t^{(j)}) \right) + L_{CL} \quad (17)$$

where $\hat{y}_t^{(j)}$ is the predicted probability of the t th frame of the j th video for the LSTM module and $\bar{y}_t^{(j)}$ is the predicted probability of the t th frame of the j th video for the TCN module.

Hybrid predictions are used to compute phase scores. The results are reported in Table 7, where we explore the triplet loss (RDCN–TCN–Triplet) and cross entropy loss (RDCN–TCN–CE) methods in (14); however, RDCN–TCN–CL performs better. With ResNet50 as a backbone, RDCN–TCN–CL outperforms all other methods with 90.2% accuracy (we retain the imagenet pretrained weights of Conv1_x to Conv4_x residual blocks in R50 while training). We also plot the bar chart (see Figure 9) of average F1 scores computed for each phase. We can observe that RDCN–TCN–CL gradually improves the F1 score performance in almost all phases compared to other baselines.

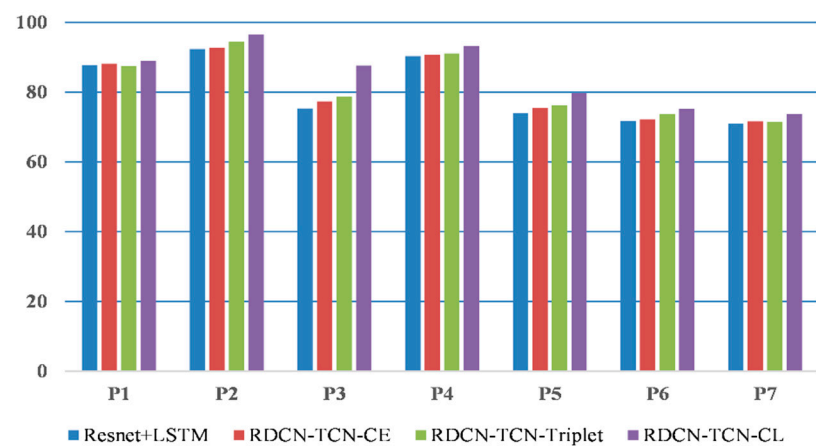


Figure 9. Comparison of the baseline on Cholec80 dataset. We report average F1 scores for each phase, P1 to P7.

7. Discussion

The automated retrieval system for the long-term archives of recorded surgeries is a key component for tasks such as surgical error analysis, surgical skills assessment, and post-operative disease analysis. In this paper, we present a content-based surgical video retrieval system (CB-SVR) to address these tasks. In contrast to previous methods for surgical video retrieval, which are based on handcrafted features, we utilize a deep learning approach to extract discriminative features from surgical videos for video retrieval. Specifically, we propose exploring a supervised contrastive learning framework under which surgical representations are learned on the top of ResNet via a hybrid LSTM-TCN module into a single end-to-end trainable network (i.e., RDCN-TCN-CL), whose activations are used to represent the surgical video. By fusing LSTM-TCN via the adaptive fusion layer, RDCN-TCN-CL ends up exploring both spatial and enhanced temporal information to represent videos. We also incorporate video frame embedding learning to further enhance the visual features needed for video representation. As the main concerns of the proposed RDCN-TCN-CL, the training setting of the network should be carefully determined after extensive consideration of system performance. To train RDCN-TCN-CL, we have shown that the length of the input video sequence affects the learning ability and therefore affects the retrieval performance. Therefore, we chose the temporal length of 24 frames for the video sequence to train the network. Additionally, increasing the length with downsampling can be investigated in future works. Meanwhile, to avoid overfitting, we also used data augmentation techniques to increase the database size. Furthermore, with extensive experiments, we demonstrated the effectiveness of RDCN-TCN-CL in end-to-end training, where RDCN-TCN-CL performs much better than its counterparts. Finally, compared to the state of the art, we demonstrated the superiority of the proposed approach for representing surgical videos in the context of video retrieval. Although RDCN-TCN-CL is slower than the state of the art, its performance is high with adequate speed. We also conduct the ablation study, where our model outperforms the state of the art in the surgical phase recognition task. With the proposed method, RDCN-TCN-CL as an attempt to exploit contrastive learning (with a deep learning approach) for video retrieval in the medical domain; we believe that it can be used to effectively analyze other medical videos, and will inspire further investigation in surgical video retrieval.

8. Conclusions

In conclusion, we present a deep learning approach for content-based surgical video retrieval. We explore the supervised contrastive learning approach to learn surgical representations. Further, to strengthen the surgical representations, we propose a hybrid approach which combines LSTM and temporal convolutions to learn temporal features under supervised contrastive learning framework. We exploit ResNet and LSTM and

TCN in a unified single joint model, i.e., RDCN–TCN–CL, to learn surgical video representations. We train the RDCN–TCN–CL under a contrastive learning approach where we explore both spatiotemporal embedding and video frame embedding learning, and the generated high-quality spatiotemporal features from the RDCN–TCN–CL model are used to retrieve similar surgical videos. With extensive experiments, we validate our proposed methodology on public surgery video datasets, where it outperforms the state of the art in retrieval and phase-recognition tasks. For retrieval, the proposed method achieves the mean average precision of 30.012 ± 1.778 vs. 22.54 ± 1.557 for Surgical Actions 160 and 81.134 ± 1.28 vs. 33.18 ± 1.311 for Cataract-101. For the phase-recognition task, the proposed method achieves 90.2% accuracy for the Cholec80 surgical dataset. For future work, we consider including training on larger database, with large batch sizes to improve the feature learning ability. Additionally, more self-supervised methods need to explore enhancing feature learning, and hashing is a worthy option that can be used for fast retrieval.

Author Contributions: Conceptualization, V.K.; methodology, V.K.; validation, B.P., S.S.A. and M.R.; formal analysis, A.G., V.K. and R.S.; writing—original draft preparation, V.K.; writing—review and editing, M.R., A.A. and A.S.A.; supervision, B.P., V.T. and A.D.; funding acquisition, S.S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Deanship of Scientific Research, Taif University Researchers Supporting Project number (TURSP-2020/215), Taif University, Taif, Saudi Arabia.

Data Availability Statement: Data in this research paper will be shared on request to the corresponding author.

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

References

1. Münzer, B.; Schoeffmann, K.; Böszörményi, L. Content-based processing and analysis of endoscopic images and videos: A survey. *Multimed. Tools Appl.* **2018**, *77*, 1323–1362. [\[CrossRef\]](#)
2. Green, J.L.; Suresh, V.; Bittar, P.; Ledbetter, L.; Mithani, S.K.; Allori, A. The Utilization of Video Technology in Surgical Education: A Systematic Review. *J. Surg. Res.* **2019**, *235*, 171–180. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Anh, N.X.; Nataraja, R.M.; Chauhan, S. Towards near real-time assessment of surgical skills: A comparison of feature extraction techniques. *Comput. Methods Programs Biomed.* **2020**, *187*, 105234. [\[CrossRef\]](#)
4. Husslein, H.; Shirreff, L.; Shore, E.M.; Lefebvre, G.G.; Grantcharov, T.P. The Generic Error Rating Tool: A Novel Approach to Assessment of Performance and Surgical Education in Gynecologic Laparoscopy. *J. Surg. Educ.* **2015**, *72*, 1259–1265. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Ritter, E.M.; Gardner, A.K.; Dunkin, B.J.; Schultz, L.; Pryor, A.D.; Feldman, L. Video-based assessment for laparoscopic fundoplication: Initial development of a robust tool for operative performance assessment. *Surg. Endosc.* **2020**, *34*, 3176–3183. [\[CrossRef\]](#) [\[PubMed\]](#)
6. van Dalen, A.S.H.M.; Legemaate, J.; Schlack, W.S.; Legemate, D.A.; Schijven, M.P. Legal perspectives on black box recording devices in the operating environment. *Br. J. Surg.* **2019**, *106*, 1433–1441. [\[CrossRef\]](#)
7. Bezemer, J.; Cope, A.; Korkiakangas, T.; Kress, G.; Murtagh, G.; Weldon, S.M.; Kneebone, R. Microanalysis of video from the operating room: An underused approach to patient safety research. *BMJ Qual. Saf.* **2017**, *7*, 583–587. [\[CrossRef\]](#)
8. Grenda, T.R.; Pradarelli, J.C.; Dimick, J.B. Using surgical video to improve technique and skill. *Ann. Surg.* **2016**, *264*, 32–33. [\[CrossRef\]](#)
9. Lavanchy, J.L.; Zindel, J.; Kirtac, K.; Twick, I.; Hosgor, E.; Candinas, D.; Beldi, G. Automation of surgical skill assessment using a three-stage machine learning algorithm. *Sci. Rep.* **2021**, *11*, 5197. [\[CrossRef\]](#)
10. Loukas, C. Video content analysis of surgical procedures. *Surg. Endosc.* **2018**, *32*, 553–568. [\[CrossRef\]](#)
11. Blum, T.; Feußner, H.; Navab, N. Modeling and segmentation of surgical workflow from laparoscopic video. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2010; Lecture Notes in Computer Science*; Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6363, pp. 400–407.
12. Lalys, F.; Riffaud, L.; Bouget, D.; Jannin, P. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 966–976. [\[CrossRef\]](#)
13. Lalys, F.; Riffaud, L.; Morandi, X.; Jannin, P. Automatic phases recognition in pituitary surgeries by microscope images classification. In *Information Processing in Computer-Assisted Interventions—IPCAI 2010; Lecture Notes in Computer Science*; Navab, N., Jannin, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6135, pp. 34–44.

14. Zia, A.; Sharma, Y.; Bettadapura, V.; Sarin, E.L.; Ploetz, T.; Clements, M.A.; Essa, I. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *Int. J. Comput. Assist. Radiol. Surg.* **2016**, *11*, 1623–1636. [[CrossRef](#)] [[PubMed](#)]
15. Weede, O.; Dittrich, F.; Worn, H.; Jensen, B.; Knoll, A.; Wilhelm, D.; Kranzfelder, M.; Schneider, A.; Feussner, H. Workflow analysis and surgical phase recognition in minimally invasive surgery. In Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012—Conference Digest, Guangzhou, China, 11–14 December 2012; pp. 1068–1074.
16. Sharma, Y.; Bettadapura, V.; Ploetz, T.; Hammerla, N.; Mellor, S.; McNaney, R.; Olivier, P.; Deshmukh, S.; McCaskie, A.; Essa, I. Video Based Assessment of OSATS Using Sequential Motion Textures. In Proceedings of the Fifth Workshop on Modeling and Monitoring of Computer Assisted Interventions (M2CAI), Boston, MA, USA, 14 September 2014; Forestier, G., Giannarou, S., Lin, H., Masamune, K., Speidel, S., Stauder, R., Penet, C., Eds.; Springer: Cham, Switzerland, 2014.
17. Allan, M.; Ourselin, S.; Thompson, S.; Hawkes, D.J.; Kelly, J.; Stoyanov, D. Toward detection and localization of instruments in minimally invasive surgery. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 1050–1058. [[CrossRef](#)]
18. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [[CrossRef](#)]
19. Shen, D.; Wu, G.; Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221–248. [[CrossRef](#)] [[PubMed](#)]
20. Pandey, B.; Pandey, D.K.; Mishra, B.P.; Rhmann, W. A Comprehensive Survey of Deep Learning in the field of Medical Imaging and Medical Natural Language Processing: Challenges and research directions. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**; *in press*. [[CrossRef](#)]
21. Blum, T.; Padoy, N.; Feußner, H.; Navab, N. Modeling and online recognition of surgical phases using hidden Markov models. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2008*; Lecture Notes in Computer Science; Metaxas, D., Axel, L., Fichtinger, G., Székely, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 5242, pp. 627–635.
22. Lalys, F.; Riffaud, L.; Morandi, X.; Jannin, P. Surgical phases detection from microscope videos by combining SVM and HMM. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging—MCV 2010*; Lecture Notes in Computer Science; Menze, B., Langs, G., Tu, Z., Criminisi, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6533, pp. 54–62.
23. Tao, L.; Elhamifar, E.; Khudanpur, S.; Hager, G.D.; Vidal, R. Sparse hidden Markov models for surgical gesture classification and skill evaluation. In *Information Processing in Computer-Assisted Interventions—IPCAI 2012*; Lecture Notes in Computer Science; Abolmaesumi, P., Joskowicz, L., Navab, N., Jannin, P., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7330, pp. 167–177.
24. Charrière, K.; Quéllec, G.; Lamard, M.; Martiano, D.; Cazuguel, G.; Coatrieux, G.; Cochener, B. Real-time analysis of cataract surgery videos using statistical models. *Multimed. Tools Appl.* **2017**, *76*, 22473–22491. [[CrossRef](#)]
25. 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 WACV, Waikoloa, HI, USA, 5–9 January 2015; pp. 1123–1129.
26. Zappella, L.; Béjar, B.; Hager, G.; Vidal, R. Surgical gesture classification from video and kinematic data. *Med. Image Anal.* **2013**, *17*, 732–745. [[CrossRef](#)]
27. Padoy, N.; Blum, T.; Ahmadi, S.A.; Feussner, H.; Berger, M.O.; Navab, N. Statistical modeling and recognition of surgical workflow. *Med. Image Anal.* **2012**, *16*, 632–641. [[CrossRef](#)]
28. Cadène, R.; Robert, T.; Thome, N.; Cord, M. M2CAI Workflow Challenge: Convolutional Neural Networks with Time Smoothing and Hidden Markov Model for Video Frames Classification. *arXiv* **2016**, arXiv:1610.05541.
29. Jalal, N.A.; Alshirbaji, T.A.; Möller, K. Evaluating convolutional neural network and hidden Markov model for recognising surgical phases in sigmoid resection. *Curr. Dir. Biomed. Eng.* **2018**, *4*, 415–418. [[CrossRef](#)]
30. Twinanda, A.P.; Shehata, S.; Mutter, D.; Marescaux, J.; De Mathelin, M.; Padoy, N. EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. *IEEE Trans. Med. Imaging* **2017**, *36*, 86–97. [[CrossRef](#)] [[PubMed](#)]
31. Al Hajj, H.; Lamard, M.; Conze, P.H.; Cochener, B.; Quéllec, G. Monitoring tool usage in surgery videos using boosted convolutional and recurrent neural networks. *Med. Image Anal.* **2018**, *47*, 203–218. [[CrossRef](#)] [[PubMed](#)]
32. Jin, Y.; Dou, Q.; Chen, H.; Yu, L.; Qin, J.; Fu, C.W.; Heng, P.A. SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Trans. Med. Imaging* **2018**, *37*, 1114–1126. [[CrossRef](#)] [[PubMed](#)]
33. Jin, Y.; Li, H.; Dou, Q.; Chen, H.; Qin, J.; Fu, C.W.; Heng, P.A. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Med. Image Anal.* **2020**, *59*, 101572. [[CrossRef](#)] [[PubMed](#)]
34. Shi, X.; Jin, Y.; Dou, Q.; Heng, P.A. LRTD: Long-range temporal dependency based active learning for surgical workflow recognition. *Int. J. Comput. Assist. Radiol. Surg.* **2020**, *15*, 1573–1584. [[CrossRef](#)] [[PubMed](#)]
35. Kreuzer, D.; Munz, M. Deep Convolutional and LSTM Networks on Multi-Channel Time Series Data for Gait Phase Recognition. *Sensors* **2021**, *21*, 789. [[CrossRef](#)]
36. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. LongTerm Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 677–691. [[CrossRef](#)]
37. Kumar, V.; Tripathi, V.; Pant, B. Learning Compact Spatio-Temporal Features for Fast Content based Video Retrieval. *Int. J. Innov. Technol. Explor. Eng.* **2019**, *9*, 2404–2409.

38. Majd, M.; Safabakhsh, R. Correlational Convolutional LSTM for human action recognition. *Neurocomputing* **2019**, *396*, 224–229. [[CrossRef](#)]
39. Li, Z.; Zhang, X.; Müller, H.; Zhang, S. Large-scale retrieval for medical image analytics: A comprehensive review. *Med. Image Anal.* **2018**, *43*, 66–84. [[CrossRef](#)] [[PubMed](#)]
40. Carlos, J.R.; Lux, M.; Giro-I-Nieto, X.; Munoz, P.; Anagnostopoulos, N. Visual information retrieval in endoscopic video archives. In Proceedings of the 2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI), Prague, Czech Republic, 10–12 June 2015; pp. 1–6.
41. Beecks, C.; Schoeffmann, K.; Lux, M.; Uysal, M.S.; Seidl, T. Endoscopic Video Retrieval: A Signature-Based Approach for Linking Endoscopic Images with Video Segments. In Proceedings of the 2015 IEEE International Symposium on Multimedia (ISM), Miami, FL, USA, 14–16 December 2015; pp. 33–38.
42. Schoeffmann, K.; Beecks, C.; Lux, M.; Uysal, M.S.; Seidl, T. Content-based retrieval in videos from laparoscopic surgery. In *Medical Imaging 2016: Image-Guided Procedures, Robotic Interventions and Modeling*; Webster, R.J., III, Yaniv, Z.R., Eds.; SPIE: Bellingham, WA, USA, 2016; Volume 9786, p. 97861V.
43. André, B.; Vercauteren, T.; Buchner, A.M.; Wallace, M.B.; Ayache, N. A smart atlas for endomicroscopy using automated video retrieval. *Med. Image Anal.* **2011**, *15*, 460–476. [[CrossRef](#)] [[PubMed](#)]
44. Beecks, C.; Kletz, S.; Schoeffmann, K. Large-Scale Endoscopic Image and Video Linking with Gradient-Based Signatures. In Proceedings of the 2017 IEEE 3rd International Conference on Multimedia Big Data (BigMM), Laguna Hills, CA, USA, 19–21 April 2017; pp. 17–21.
45. Droueche, Z.; Lamard, M.; Cazuguel, G.; Quellec, G.; Roux, C.; Cochener, B. Motion-based video retrieval with application to computer-assisted retinal surgery. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), San Diego, CA, USA, 28 August–1 September 2012; pp. 4962–4965.
46. Quellec, G.; Lamard, M.; Droueche, Z.; Cochener, B.; Roux, C.; Cazuguel, G. A polynomial model of surgical gestures for real-time retrieval of surgery videos. In *Medical Content-Based Retrieval for Clinical Decision Support—MCBR-CDS 2012; Lecture Notes in Computer Science*; Greenspan, H., Müller, H., Syeda-Mahmood, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7723, pp. 10–20.
47. Syeda-Mahmood, T.; Ponceleon, D.; Yang, J. Validating cardiac echo diagnosis through video similarity. In Proceedings of the 13th ACM International Conference on Multimedia (MM), Singapore, 6–11 November 2005; pp. 527–530.
48. Quellec, G.; Charrière, K.; Lamard, M.; Droueche, Z.; Roux, C.; Cochener, B.; Cazuguel, G. Real-time recognition of surgical tasks in eye surgery videos. *Med. Image Anal.* **2014**, *18*, 579–590. [[CrossRef](#)] [[PubMed](#)]
49. Quellec, G.; Lamard, M.; Cazuguel, G.; Droueche, Z.; Roux, C.; Cochener, B. Real-time retrieval of similar videos with application to computer-aided retinal surgery. In Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Boston, MA, USA, 30 August 3–September 2011; pp. 4465–4468.
50. Droueche, Z.; Lamard, M.; Cazuguel, G.; Quellec, G.; Roux, C.; Cochener, B. Content-based medical video retrieval based on region motion trajectories. In Proceedings of the 5th European Conference of the International Federation for Medical and Biological Engineering, Budapest, Hungary, 14–18 September 2011; Jobbágy, Á., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; Volume 37, pp. 622–625.
51. Muenzer, B.; Primus, M.J.; Kletz, S.; Petscharnig, S.; Schoeffmann, K. Static vs. Dynamic Content Descriptors for Video Retrieval in Laparoscopy. In Proceedings of the 2017 IEEE International Symposium on Multimedia (ISM), Taichung, Taiwan, 11–13 December 2017; pp. 216–223.
52. Kletz, S.; Schoeffmann, K.; Munzer, B.; Primus, M.J.; Husslein, H. Surgical action retrieval for assisting video review of laparoscopic skills. In Proceedings of the MultiEdTech 2017—Proceedings of the 2017 ACM Workshop on Multimedia-Based Educational and Knowledge Technologies for Personalized and Social Online Training, Co-Located with MM 2017, Mountain View, CA, USA, 27 October 2017; pp. 11–19.
53. Amanat, S.; Idrees, M.; Khan, M.U.G.; Rehman, Z.; Chang, H.; Mehmood, I.; Baik, S.W. Video retrieval system for meniscal surgery to improve health care services. *J. Sens.* **2018**, *2018*, 4390703. [[CrossRef](#)]
54. Schoeffmann, K.; Husslein, H.; Kletz, S.; Petscharnig, S.; Muenzer, B.; Beecks, C. Video retrieval in laparoscopic video recordings with dynamic content descriptors. *Multimed. Tools Appl.* **2018**, *77*, 16813–16832. [[CrossRef](#)]
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
56. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
57. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*; Curran Associates, Inc.: Red Hook, NY, USA; pp. 3104–3112.
58. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1003–1012. [[CrossRef](#)]
59. Czempel, T.; Paschali, M.; Keicher, M.; Simson, W.; Feussner, H.; Kim, S.T.; Navab, N. TeCNO: Surgical Phase Recognition with Multi-stage Temporal Convolutional Networks. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020; Lecture Notes in Computer Science*; Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, K.S., Racocanu, D., Joskowicz, L., Eds.; Springer: Cham, Switzerland, 2020; Volume 12263, pp. 343–352. [[CrossRef](#)]

60. Ramesh, S.; Dall'Alba, D.; Gonzalez, C.; Yu, T.; Mascagni, P.; Mutter, D.; Padoy, N. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 1111–1119. [[CrossRef](#)]
61. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (ICML), Vienna, Austria, 12–18 July 2020; pp. 1597–1607.
62. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
63. Saxe, A.M.; McClelland, J.L.; Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
64. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
65. Schoeffmann, K.; Taschwer, M.; Sarny, S.; Münzer, B.; Primus, M.J.; Putzgruber, D. Cataract-101—Video dataset of 101 cataract surgeries. In Proceedings of the 9th ACM Multimedia Systems Conference (MMSys), Amsterdam, The Netherlands, 12–15 June 2018; pp. 421–425.
66. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 27*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; pp. 1097–1105.
67. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9.
68. Twinanda, P. Vision-Based Approaches for Surgical Activity Recognition Using Laparoscopic and RGBD Videos. Ph.D. Thesis, Université de Strasbourg, Strasbourg, France, 2017.