Learning spatiotemporal features for esophageal abnormality detection from endoscopic videos

Learning spatiotemporal features for esophageal abnormality detection from endoscopic videos

Noha Ghatwary, Massoud Zolgharni, Faraz Janan and Xujiong Ye

***Abstract*— Esophageal cancer is categorized as a type of disease with a high mortality rate. Early detection of esophageal abnormalities (*i.e. precancerous and early cancerous*) can improve the survival rate of the patients. Recent deep learning-based methods for selected types of esophageal abnormality detection from endoscopic images have been proposed. However, no methods have been introduced in the literature to cover the detection from endoscopic videos, detection from challenging frames and detection of more than one esophageal abnormality type. In this paper, we present an efficient method to automatically detect different types of esophageal abnormalities from endoscopic videos. We propose a novel 3D Sequential DenseConvLstm network that extracts spatiotemporal features from the input video. Our network incorporates 3D Convolutional Neural Network (3DCNN) and Convolutional Lstm (*ConvLstm*) to efficiently learn short and long term spatiotemporal features. The generated feature map is utilized by a region proposal network and ROI pooling layer to produce a bounding box that detects abnormality regions in each frame throughout the video. Finally, we investigate a post-processing method named Frame Search Conditional Random Field (FS-CRF) that improves the overall performance of the model by recovering the missing regions in neighborhood frames within the same clip. We extensively validate our model on an endoscopic video dataset that includes a variety of esophageal abnormalities. Our model achieved high performance using different evaluation metrics showing 93.7% recall, 92.7% precision, and 93.2% F-measure. Moreover, as no results have been reported in the literature for the esophageal abnormality detection from endoscopic videos, to validate the robustness of our model, we have tested the model on a publicly available colonoscopy video dataset, achieving the polyp detection performance in a recall of 81.18%, precision of 96.45% and F-measure 88.16%, compared to the state-of-the-art results of 78.84% recall, 90.51% precision and 84.27% F-measure using the same dataset. This demonstrates that the proposed method can be adapted to different gastrointestinal endoscopic video applications with a promising performance.**

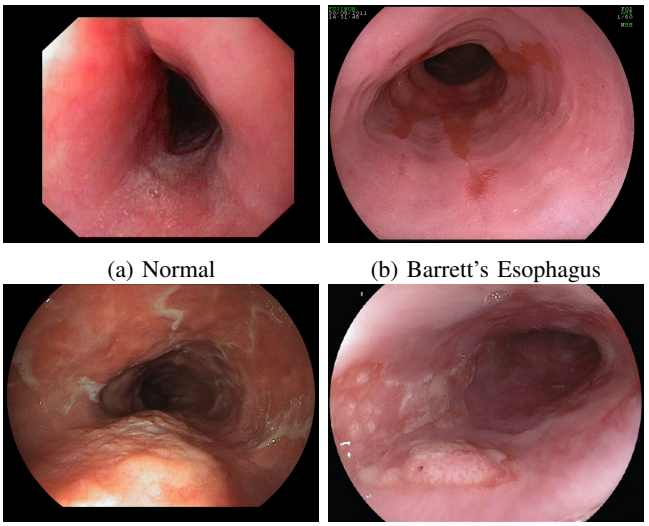
***Index Terms*— Esophageal abnormality detection, endoscopy, spatio-temporal features, deep learning.**

This paragraph of the first footnote will contain the date on which you submitted your paper for review.

N. Ghatwary, F. Janan and X. Ye are with the Computer Science Department at the University of Lincoln, Lincoln, UK (e-mail: nghatwary@lincoln.ac.uk, fjanan@lincoln.ac.uk, xye@lincoln.ac.uk).

N. Ghatwary is with the Computer Engineering Department at the Arab Academy for Science and Technology, Alexandria, Egypt (e-mail: noha.ghatwary@aast.edu).

M. Zolgharni is with the Computer Science Department at the University of West London, London, UK (e-mail: massoud.zolgharni@uwl.ac.uk).



(a) Normal (b) Barrett's Esophagus

(c) Esophageal Adenocarcinoma (d) Squamous Cell Carcinoma

Fig. 1: Example of different tissues deformation cases of the esophagus: (a) Normal, (b) Precancrous (**BE**), (c) Cancerous (**EAC**), (d) Cancerous (**SCC**).

I. Introduction

ESOPHAGEAL cancer (EC) is ranked the sixth leading cause of death from cancer and the seventh most common cancer in the world [1]. Based on the progression of the EC, the survival rate can vary from 5% to 45% based on the stage and the development of the disease giving an average of 19% survival rate on a 5-year overall plan [2], [3]. Early detection and treatment are important in decreasing the mortality rate and increasing the survival rate [4]. EC is an aggressive type of cancer that has no symptoms until the late stages. There exists two types of EC: *Esophageal Adenocarcinoma (EAC)* that takes place in gland cells and *Squamous Cell Carcinoma (SCC)* that appears in squamous cells. Most instances arise from the undetected precancerous abnormalities such as *Barrett's Esophagus* (BE). BE is the process of growing abnormal cells (i.e. metaplastic intestinal epithelium) that replace the normal cells in the esophageal area [5]. Moreover, BE is counted as the main esophageal precancerous condition that needs to be detected in the early stages to avoid evolving into cancer [6]. Fig. 1 illustrates examples of endoscopic view for Normal, BE, EAC & SCC cases.

Endoscopy is the tool used for the detection of esophageal abnormalities (including precancerous and cancerous) [7]. Although the endoscopy has been available for a couple of

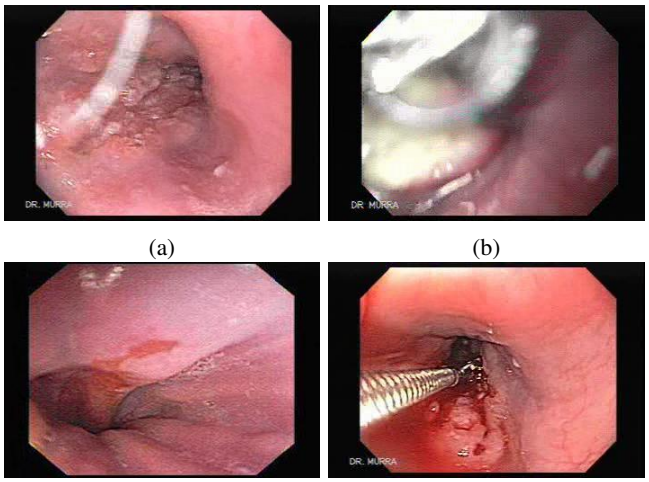


Fig. 2: Examples of challenges frames from esophageal endoscopic videos. (a) Low-quality image, (b) blurred image, (c) challenging appearance, (d) Tool appearance.

decades for the clinical examination, the process of detection is considered challenging for several reasons. First, the esophageal abnormalities have different sizes and shapes that can be located randomly throughout the esophagus tube which makes it challenging to capture [8]. Secondly, the diagnose of BE is sometimes missed due to the failure in discriminating between the columnar mucosa from metaplastic epithelium [9]. Additionally, the process of automatic detection from videos is an extremely challenging task. The rapid movement of the endoscope produces low-quality and blurry images (Fig. 2a & Fig. 2b). Moreover, the endoscope is not always centered on the examined region with a limited abnormality view inside the esophagus (Fig. 2c). Furthermore, the occurrence of intestinal juices and tool appearance can block the presence of the abnormality (Fig. 2d). Therefore, detecting abnormalities from the esophageal endoscopic videos is very different from detecting the abnormalities from selected images.

All the previous work put much effort into studying automatic detection of esophageal abnormalities from selected still frames/ images from videos (i.e. as will be discussed in next section). Nevertheless, No work in the literature has focused on finding the different types of esophageal abnormalities from videos or a sequence of frames. Whist, deep learning methods have been investigated for the automatic detection of polyps from colonoscopic videos (i.e. lower GI tract) [10]. Recent studies investigated combining temporal information with spatial information from colonoscopy videos for more accurate polyp detection [11]. A two-step approach was introduced by [12], the method first extracts geometric features: *color, texture clues* and *shape context* to detect candidate regions. The spatiotemporal pattern appearance of the polyps are learned by extracting CNN from three consecutive frames separately for each candidate region using 2d-CNN. Later studies confirmed that the 3D network designs are more suitable for the video datasets [13]. Yu et al. presented 3D fully convolutional network (3DFCN) to detect polyps in colonoscopic videos. The 3D-CNN extracts spatiotemporal features by performing convolutions along the video. The evaluation results for this model showed the capability of 3DCNN to learn more illustrative spatiotemporal features from colonoscopic videos compared to 2D networks. Moreover, other methods made use of temporal dependencies between a consecutive set of video frames to provide useful information in detecting polyps when combined with spatial information [15], [16].

Furthermore, extracting spatiotemporal features from videos using 3D-CNN has proved its effectiveness in the non-medical field such as action [17] and object recognition [18]. Tran *et. al* [19] showed that extracting spatiotemporal features using 3DCNNs to represent the model appearance and motion simulation in videos are more suitable than learning spatial features only using 2D ConvNets. Zhou *et. al* [20] suggested a video copy detection system utilizing spatiotemporal features by dividing the video sequence into multiple video clips and extracting CNN features from the spatial and temporal domains of each video clip. Results showed that using the suggested spatiotemporal CNN features gain a high performance in terms of both accuracy and effectiveness. Moreover, Colleoni *et. al* [21] proposed a 3DCNN architecture for surgical-instrument and joint-connection detection from videos by extracting spatiotemporal features that effectively increased the segmentation performance compared to 2D models using features from single-frames.

In computer-aided diagnosis, the high precision and recall results are important to provide accurate detection of abnormality. In the methods presented by [12]–[16], the extracted spatiotemporal features showed its effectiveness when incorporated in the model by producing a high precision value, but a relatively low recall value was shown that needs to be improved. Therefore, the enhancement of extracting distinctive spatiotemporal features with an appropriate model should be investigated to have a better overall detection performance.

Recently, Convolutional Long Short-Term Memory (ConvLstm) (i.e. type of Recurrent Neural Network (RNN)) showed its capability to encode long-term temporal information [22]. The ConvLstm can learn the spatiotemporal consistency across the surgical video frames [23] and preserve the spatiotemporal regularity between neighboring frames [24]. Studies proved the efficiency of ConvLstm in learning the temporal variable characteristics from the sequence of frames when incorporated in deep learning networks [25]. When the ConvLstm is included with 3D-CNN, the network is able to extract the short-term and long-term temporal information along with spatial information; producing a feature map that covers the spatiotemporal features of a longer sequence of video frames [26], [27].

In this paper, we propose a novel 3D Sequential Dense-Convlstm Faster R-CNN for the detection of esophageal abnormalities (cancerous and precancerous) from endoscopic videos. The network is built using the concept of the densely connected convolutional network (DenseNet) [28], which propagates the gradient and feature information throughout the network by taking each layer as input for all its upcoming layers. In our model, the DenseNet has been modified in several aspects where we propose increasing the internal

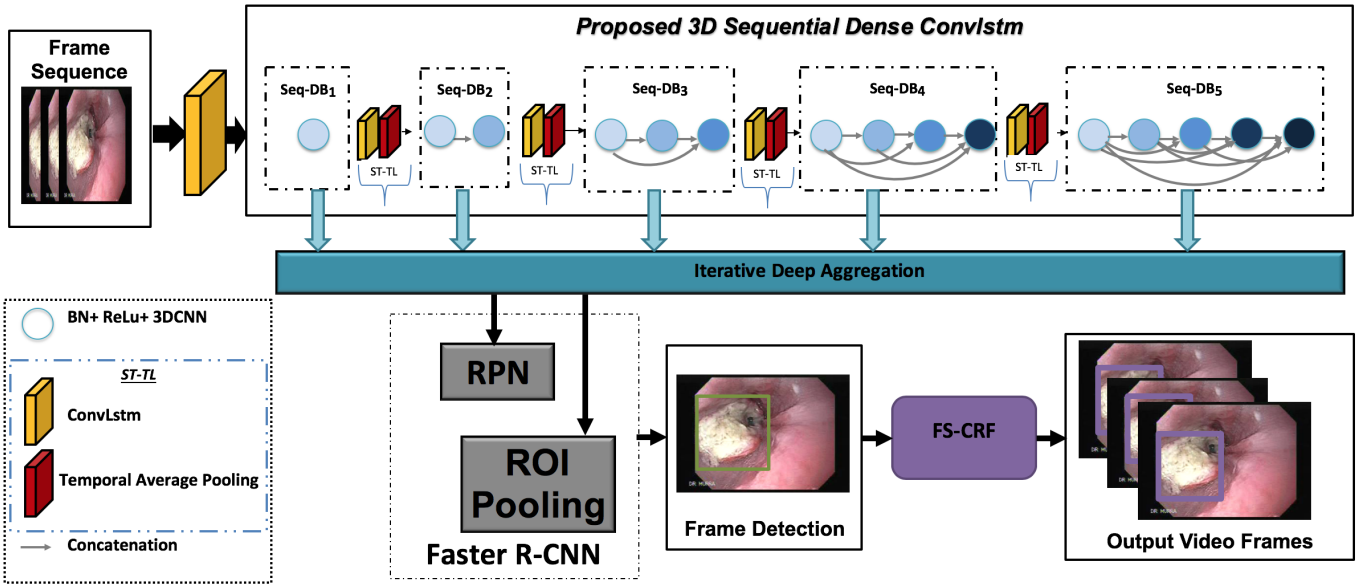


Fig. 3: Overview of the abnormality detection approach. First, Spatiotemporal features are extracted from the video input using the proposed 3D DenseConvLstm network. Then the features are used by Faster R-CNN to generate BBs for EAC regions in the video. Finally, the proposed FS-CRF post-processing is applied to improve the general detection performance.

layers of the dense blocks in the network sequentially to provide more related information. Moreover, for the construction of the network, we utilize the 3D-CNN with ConvLstm [22] to extract spatiotemporal features. The 3D-CNN extracts features regarding the third dimension (i.e. *Time*) holding richer information while the ConvLstm explores the relation of spatiotemporal information between video frames. The architecture of the proposed network is designed to extract features from videos and allows each frame to learn features from subsequent frames. Moreover, the extracted features are then used by the Faster R-CNN to generate bounding boxes to locate the abnormalities throughout the video.

To improve the overall detection performance, we propose a post-processing method named Frame Search Conditional Random Field (FS-CRF). The proposed FS-CRF employs a frame search algorithm with the Dense CRF on a frame-based level to improve the performance by recovering missed regions and removing false positives. To validate the efficiency of our proposed method, we evaluate it on a large dataset that is composed of different types of abnormalities. The main contribution of our work is four-folds:

- An effective approach for the detection of different types of esophageal abnormalities (BE, EAC and SCC) from endoscopic videos is presented. To the best of our knowledge, this is the first first study of its kind to detect different types of esophageal abnormalities using the full videos instead of selected frames.
- We design a novel 3D Sequential Dense-Convlstm backbone network to extract features from esophageal endoscopic videos. By incorporating the 3D-CNN and the ConvLstm, the proposed network has the ability to learn spatiotemporal features that is more compatible with the properties of videos.
- We implement an FS-CRF for post-processing that can recover the missed abnormal regions in a sequence of consecutive frames based on the initial detection output to improve the overall detection performance.
- We extensively validate the proposed model using endoscopic videos dataset that includes normal, precancerous and cancerous patients. Moreover, we compare the performance of the model with different types of networks such as the 2D-CNN network and U-Net.

The remainder of this paper is organized as follows: Section II, provides an overview of the deep learning esophageal abnormality state-of-the-art methods. In Section III the proposed 3D Sequential DenseConvLstm model is explained in detail. Section IV describes the dataset used for evaluation and Section V reports the experimental results with the discussion. Finally, we conclude this work in Section VI.

II. RELATED WORK

In the literature, methods for automatic detection of esophageal abnormalities are divided into two categories: Handcrafted features based methods and CNN based methods. This section briefly reviews **only** deep learning methods that automatically detects esophageal EAC (i.e. only publicly available dataset by MICCAI'15 [29]) regions from endoscopic images. More details about other methods and using different examination modalities are discussed in [9], [30], [31].

Mendel *et. al* [32], detected EAC regions by extracting CNN features from patches of endoscopic images. A 50-layer deep residual network (ResNet) was initialized with ImageNet parameters to learn features from the images through transfer learning. Later, Reil et *et. al* [33], also used transfer learning with ImageNet parameters to assess different architectures such as *VGG'16, GoogleNet, AlexNet* to extract features from

the images and detect EAC regions. The extracted features were classified using conventional classifiers (i.e. SVM & RF). Furthermore, in our previous work [34], we evaluated different deep learning methods such *Regional-Based Convolution Neural Network (R-CNN) [35], Fast R-CNN [36], Faster R-CNN [37] and Single Shot Detector (SSD) [38]*. Each of these deep learning methods used the CNN features extracted by the VGG'16 network to generate bounding boxes that locate EAC regions. Afterward, in [39], handcrafted Gabor Filter responses were obtained from endoscopic images and merged with CNN features extracted from DenseNet. The combined features were used by the Faster R-CNN to locate two types of esophageal abnormalities: *Esophgities* and *EAC* from HD-WLE images. Later on, a Gabor Fractal (GF) image was generated from the different Gabor filter responses of endoscopic images in [40] to create a two-input network for esophageal abnormality detection. CNN features extracted from the original input image were used to generate region proposals of abnormal candidates use the RPN network in Faster R-CNN. Bilinear fusion is then employed to fuse features from both the original and GF image to detect the final abnormal region based on suggested proposals.

The aforementioned methods focused on finding abnormal regions (mainly EAC) from selected images. To the best of our knowledge, no work in the literature has focused on finding the different types of esophageal abnormalities from videos and challenging frames.

III. Method

In Fig. 3 we illustrate the proposed automatic detection framework which involves three main stages: *(a) spatiotemporal feature extraction, (b) detection of abnormality regions* and *(c) post-processing phase*. As shown, first the spatiotemporal features are extracted using a novel 3D Sequential DenseConvLstm Network that is equipped with dense connectivity and integrated with both 3DCNN and ConvLstm. The integration between the feature extracted from 3DCNN and ConvLstm preserves the global temporal connectivity between subsequent frames. Moreover, we set a ConvLstm layer to be the initial filter for the video input. Afterwards, the extracted features from each dense block are aggregated together. The extracted spatiotemporal features are used by the region proposal network (RPN) and region-of-interest pooling layer (ROI Layer) in the Faster R-CNN to detect the region of abnormality in each video frame. Finally, the detection results are post-processed with a proposed FS-CRF to improve the final performance of the model. In the remainder of this section, we explain each stage with its components in detail.

A. Spatiotemporal Feature Extraction:

1) 3D Sequential DenseConvLstm: The network is built on the concept of DenseNet architecture [28], that encourages feature reuse by connecting the output of a layer to all upcoming layers in the network. The proposed 3D Sequential DenseConvLstm is made up of three main components: *Sequential Dense Block*, *SpatioTemporal Transition Layer* and *Growth Rate*.

- **Sequential Dense Block(Seq-DB)**
  The **DB** performs the operation of the dense connectivity where each layer takes the feature map of all previous layers as an input. The output of a DB is given by:

$$x_l = H_l([x_0, x_1, ..., x_{l-1}]) \quad (1)$$

  where x_l is the feature map output from each DB with l layers. The $H(.)$ represents the composite operation formed of Batch Normalization (BN), Relu, and Convolutional operation. The (.) denotes the concatenation process of the feature maps inside each DB block where each feature map has the same size (i.e. height and width). For the **Seq-DB** we propose two main contributions; first, we increase the number of internal layers sequentially, namely, the number of layers (l) per block is equal to the DB position (N) in the network (as shown in 3). For Simplicity, the output for each block Seq-DB$_N$ is x_N where N is the DB block number (i.e. position) in the network (i.e. $x_N = H_N[x_0, x_1, ..., x_{N-1}]$). The DenseNet depends more on high-level features than low-level ones, therefore, increasing the features in later DB's will provide more global features [41]. Accordingly, it can maintain high performance with a reduced number of trained parameters. Secondly, we propose using 3DCNN to learn local spatiotemporal features instead of 2D CNN used in the original densenet. The 3DCNN has the ability to extract short-term temporal features along with the spatial information, therefore, it is useful to use with the video analysis [42]. For the 3DCNN, the operation convolves a 3D filter therefore both the feature map and kernel have a depth dimension (i.e. spatial and temporal as depth). The operation of each layer (l) in Seq-DB is: (BN, Relu, & (3×3×3) 3DCNN).
- **SpatioTemporal-Transition Layer (ST-TL)** Furthermore, the **ST-TL** exists between each DB which helps downsample the feature map. In the network, the proposed ST-TL is composed of (1×1×1) *ConvLstm* with stride=1 × 1 × 1 and same-padding to maintain the size of spatiotemporal feature map. Also, a (2×2×2) temporal average pooling is applied after the ConvLstm layer.
  - **Convolution Lstm (ConvLstm) Unit:** ConvLstm is a type of Recurrent neural networks (RNNs). RNNs proved the ability to learn temporal information in several fields [43]. Precisely, ConvLstm is a development of the LSTM with a convolution operation inside the LSTM cell. ConvLstm uses convolution operation instead of matrix multiplication at each gate of the LSTM cell. The ConvLSTM is designed for 3-D data as its input (i.e. such as videos) while LSTM input data are one-dimensional.
    The ConvLstm is capable of learning long-term spatiotemporal information by encoding the changes of spatial and temporal information using the convolution gates. Moreover, it forces consistency across time by taking into account the features from previous frames leading to an improved detection. The operations of convolution and recurrence of the input-

to-state and state-to-state transitions benefit from the spatiotemporal correlation information. In our model we use the original ConvLstm as explained in [22]. Additionally, in our model, the ConvLstm layer is also used as the initial filter applied to video input before the 3D Sequential DenseConvLstm network. The ConvLstm is initially used to set the number of frames to learn long-term spatiotemporal information for the current frame.

– **Temporal Average Pooling:** Despite the fact that the ConvLstm efficiently extract spatiotemporal features but it might be biased towards the end frames in the sequnce which can reduce the efficiency in extracting appropriate information over the full sequence. Therefore, we utilize the *temporal average pooling* to capture long-term features present by considering information through the sequence and it downsamples the feature map size .

- **Growth Rate** Each Seq-DB produces a feature map of size f (i.e. generated by eq.1) that is controlled by the growth rate (G) [28]. The G is a small integer value that regulates the amount of new information held by each layer. Moreover, it controls the growth of the network and improves parameters efficiency. Therefore, the size of L^{th} layer is $f * (l-1) + f_0$ where f_0 represents the size of initial filter.

2) Iterative Deep Aggregation: Finally, as shown in Fig. 3, we aggregate the features extracted from each **Seq-DB**$_N$ to produce the final feature map through *iterative deep aggregation* [44]. The process of aggregation starts with the shallow layers and then iteratively merges with deeper layers. The aggregation between deep and shallow layers has been demoestrated to improve the overall performance of the network with high-resolution feature map [45]. The process of aggregation starts with the first shallow layer and keep merging with deeper layers iteratively as follows:

$$F(d_1, ..., d_n) = \begin{cases} d_1 & if \quad n = 1 \\ F(A(d_1, d_2), ..., d_n) & otherwise \end{cases} \tag{2}$$

where, d_n represent the features extracted from Seq-DB$_N$ for n=number of the Seq-DB blocks. And, A represents the aggregation node and F final feature map used by the next phase for detection. To combine the features from different layers, we downsample the low-level feature through convolution and merge it with the following level features.

B. Detection of Abnormality Region: Faster R-CNN

In our model, the feature map produced by the 3D Sequential DenseConvLstm is employed by the Faster R-CNN to generate bounding boxes (BBs) that detect abnormal esophageal regions in the endoscopic video frames. The Faster R-CNN is composed of two main stages: *RPN* and *ROI pooling layer* that share the same feature map to reduce computational complexity. The RPN is responsible for generating a list of candidate BBs with a confidence score that might hold an abnormal region. The RPN generates anchor boxes that have different sizes and scales to provide K proposals for each location. For each frame, there exists $(W \times H \times K)$ proposals, where W and H represent the size of the feature map. The RPN network has two output layers; the first is a classifier layer that produces a probability whether the proposed anchor box contains an abnormality or not. The other layer is a regression layer that adjusts the high probability boxes to better fit the detected object. The boxes with the highest score are called region proposals and they are sent to the next phase. During the training phase the classification and regression outputs from the RPN proposals rely on an Intersection-Over-Union (IoU) threshold which is used to measure the ratio of the overlapping and union area between the ground truth and the predicted bounding box area.

Afterwards, the ROI pooling unifies the feature map of each proposal generated by the RPN layer and classifies them using softmax into normal, precancerous and cancerous regions. Features in this phase are reused from the same feature map used by the RPN layer as they both share the same convolution layer. Moreover, the ROI pooling has a regression layer that produces the coordinates of the BB (c_x, c_y, w, h) that locates the detected region. Further details about Faster R-CNN can be found in the original paper [37].

C. Post-Processing Phase: Frame Search Conditional Random Field (FS-CRF)

The target of the post-processing stage is to improve the overall detection performance of the model by removing false positives (FPs) and recovering missing abnormal regions in a sequence of frames obtained from the previous step. The proposed FS-CRF is constructed on two stages; a *frame search* algorithm and *densely CRF* applied on a frame base level. The pseudocode for the proposed FS-CRF post-processing method is summarized in Algorithm 1. Each stage is described in detail below.

1) Frame Search algorithm: The *frame search* algorithm has two main functions (i) to remove false positives (FPs), and (ii) to recover missing regions, by searching for its nearest labeled frames (i.e. frames with detection).

- *Removing FPs:* For any frame (f) with **detection**, the algorithm searches if there exists another detection within a frame-window threshold t (for example, if $t = 7$ then $f +/- 7$). If the current frames is the only detection then this frame is counted as FP and the detected BBs are removed.
- *Nearest Labeled Frames (L_x, L_y)*: For any frame (L_f) with **no detection**, the algorithm search for the nearest frames with detection labels named L_x and L_y within the window frame t. Then it checks if the Intersection over Union (IoU) of the BBs in these frames as follows:

$$IoU = \frac{B_x \cap B_y}{B_x \cup B_y} \tag{3}$$

Where, B_x and B_y represent the area of the generated BB for two nearest frames (L_x, L_y) respectively. If the IoU is greater than 0.7 then these frames are considered to have the same abnormal region. A labeled image (L) is generated from the intersection of the labels of L_x and

L_y (as shown in Fig. 4(c)). The label image (L) is then used in the next stage to find missed regions along with frames L_x and L_y .

2) CRF: We adopt the densely CRF proposed by [46] and apply it on a frame base level using the labeled frame (L) generated from the first stage to find missed abnormal regions in frame (L_f) (i.e. frame with no detection). For an input frame (L_f) and label (L), the CRF energy function is given by:

$$E(l) = \sum_i \sigma_u(l_i) + \sum_{i<j,h} \sigma_p(l_i, l_j, l_h) \tag{4}$$

where the unary potential is defined as the negative log-likelihood $\sigma_u(l_i) = -\log Q(l_i|L_f)$ that measure the energy cost of assigning the label (l_i) to pixel i in frame L_f. Where, $Q(l_i|L_f)$ is obtained from the output of Faster R-CNN using the proposed 3D Sequential DenseConvLstm for nearest labeled frames (L_x, L_y). The pairwise potential $\sigma_p(l_i, l_j, l_h)$ is defined as a linear combination Gaussian kernels (where i, j and h are pixels from frames L_f, L_x, and L_y respectively) given by:

$$\begin{aligned}\sigma_p(l_i, l_j, l_h) = \mu(l_i, l_j) \underbrace{\sum_{n=1}^{N} w^n k^n(z_{i,f}, z_{j,x})}_{k(z_{i,f}, z_{j,x})} \\ + \mu(l_i, l_h) \underbrace{\sum_{n=1}^{N} w^n k^n(z_{i,f}, z_{h,y})}_{k(z_{i,f}, z_{h,y})}\end{aligned} \tag{5}$$

where w^n is the linear combination weight, z_i, z_j and z_h are feature vectors for pixels i, j and h in an arbitrary feature space, μ represents the label compatibility function and k^n for $n = 1, 2, ..., N$ representing the Gaussian kernels. Following [46], we use **two** kernels defined as: $k(z_{i,f}, z_{j,x}) = w_1 \exp(-\frac{|p_{i,f}-p_{j,x}|^2}{2\delta_\alpha^2} - \frac{|I_{i,f}-I_{j,x}|^2}{2\delta_\beta^2}) + w_2 \exp(-\frac{|p_{i,f}-p_{j,x}|^2}{2\delta_\gamma^2})$ and $k(z_{i,f}, z_{h,y}) = w_1 \exp(-\frac{|p_{i,f}-p_{h,y}|^2}{2\delta_\alpha^2} - \frac{|I_{i,f}-I_{h,y}|^2}{2\delta_\beta^2}) + w_2 \exp(-\frac{|p_{i,f}-p_{h,y}|^2}{2\delta_\gamma^2})$. The first term in k relies on pixel position (p) and color (I) (appearance kernel) and the second term depend on (p) only (smoothness kernel). The Gaussian kernel is controlled by the scale (δ).

The target is to minimize the value E(l) of the CRF energy function to produce the most probable label for each pixel in L_f. The energy function is approximately estimated by using the mean-field inference algorithm [46] to compute the distribution $Q(L)$ to create the new label (L). As summarized in Algorithm 2, the approximation of $Q(L)$ is optimized iteratively by applying a sequence of message passing that updates a single variable through incorporating information from the variables of the nearest two frames L_x and L_y. Using the output Label image (CRF_{Label}) from the CRF we generate a bounding box for the unlabeled frame (f) (as shown in Fig. 4(d) and 4(e)).

Fig. 4 displays samples of the output after applying the CRF to the unlabelled frame. The figure presents two examples: i) a frame with generated detection and ii) a frame with no detection. This shows that using CRF does not always generate bounding boxes for each unlabeled frame which proves the robustness of the post-processing phase.

Algorithm 1 Proposed Frame Search Algorithm Steps Description

Input: Video frames with BBs from Faster R-CNN using 3D Sequential DenseConvLstm

for $f = 1$ **to** N **do** { N : no. Video frames}
  $counter = 0$
  $Label =$ label for detection of frame f
  if $label$ **then** {Has a Label}
    for $i = f + 1$ **to** t **do** {t=frame threshold}
      if $label$ **then**
        $counter++$
      end if
    end for
    for $i = f - t$ **to** $f - 1$ **do** {t=frame threshold}
      if $label$ **then**
        $counter++$
      end if
    end for
    if $counter <= 1$ **then**
      Remove Label (considered as FP)
    end if
  else {Has **no** Label}
    for $i = f + 1$ **to** t **do** {t=frame threshold}
      if $label$ **then**
        $counter++$
        $L_x \leftarrow f$ {frame with nearest label before}
      end if
    end for
    for $i = f - t$ **to** $f - 1$ **do** {t=frame threshold}
      if $label$ **then**
        $counter++$
        $L_y \leftarrow f$ {frame with nearest label after}
      end if
    end for
    $IoU_{BB_{xy}} = IoU(BB[L_x], BB[L_y])$
    if $counter >= 2$ **and** $IoU_{BB_{xy}} > 0.7$ **then**
      $L \leftarrow interstion(BB[L_x], BB[L_y]))$
      $L_f \leftarrow$ frame (f) with no label
      $CRF_{Label} \leftarrow CRF(L_f, L_x, L_y, L)$
      Generate BB from CRF_Label for L_f
    end if
  end if
end for

Output: Updated video frames BBs from FS-CRF

IV. Materials

In this study, we carried out the experiments using a dataset of videos obtained from the open-access website Gastrointestinal Atlas [47]. The Gastrointestinal Atlas provides a large high-resolution video dataset for gastrointestinal endoscopy. We only selected the videos concerning the esophagus for our experiments. The dataset includes 44 endoscopic videos

Algorithm 2 Mean Field Algorithm for Proposed Frame-Based CRF

Input: L_f, L_x, L_y and L
Output: Q

$Q \leftarrow \frac{1}{Z_i} exp\{-\sigma_u(l_i)\}$
while not converged **do**
  $\tilde{Q_i}^n(l_i) \leftarrow (\sum_{\forall i \neq j} k^n(z_{i,f}, z_{j,x}) Q_j(l_i) + \sum_{\forall i \neq h} k^n(z_{i,f}, z_{h,y}) Q_k(l_i))$
  $\hat{Q}_i(l_i) \leftarrow \sum_{l \in L} \mu^n(L_{i,f}, l_i) \sum_n w^n \tilde{Q_i}^n(l_i)$
  $Q_i(l_i) \leftarrow exp\{-\sigma_u(l_i) - \hat{Q}_i(l_i)\}$
  normalize $Q_i(l_i)$
end while

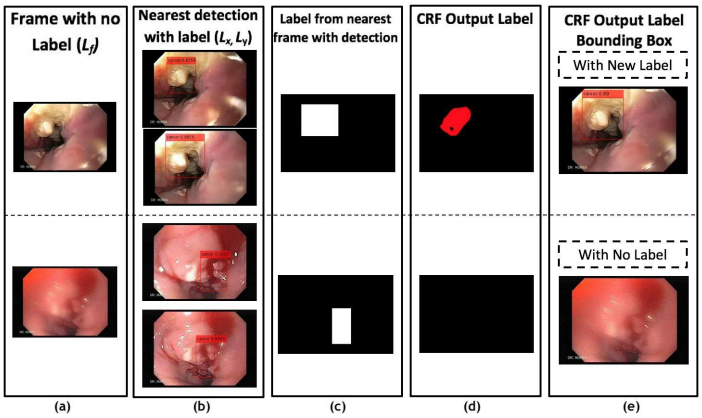


Fig. 4: Examples of the generated bounding-box using the CRF to find the abnormal region in unlabelled frame L_f using the label from the nearest labeled frames L_x & L_y. The first row displays an example of prediction after FS-CRF post-processing, while the second row represents an example of no prediction.

(total of 42,425 frames) involving four different sub-classes of three categories: *Normal, Precancerous: BE, Cancerous: EAC and SCC*. Each video has an average duration of 50 seconds (the time ranges from 30 seconds to 4 minutes per video). Additionally, the videos have a resolution of 240×352 pixels with a frame rate of 30 frames per second (fps) and divided into three categories; normal, precancerous and cancerous. The 44 videos are classified as follows: *Normal* (1 video), *BE* (24 videos), *EAC* (9 videos) and *SCC* (10 videos). The abnormal regions in the dataset were manually annotated by experts in the field. These annotations are used as ground-truth. The dataset was randomly divided into 50% training, 20% validation, and 30% testing, which is also based on the abnormality type to make sure that videos from the same patient are only present in either training, validation or testing set (i.e. to avoid any bias/overfitting results). Samples frames from the video dataset with the annotations are illustrated in Fig. 5.

V. Experiments and Results

In this section, experiments are carried out to evaluate the performance of the proposed method using the dataset described in the previous section (Sec. IV). The implementation details, parameter setting of the models, and evaluation protocols are described. Comprehensive experiments have been carried out to demonstrate the advantages of our proposed method by comparing with different methods, including with and without the proposed FS-CRF post-processing phase, using 2D model and our 3D Sequential DenseConvLstm network. Experimental results have been provided with illustrative examples from the detection output, and further discussions in terms of quantitative and qualitative evaluations have also been provided. To justify the design of the 3D Sequential DenseConvLstm network, we present a series of experiments to demonstrate the impact of each contribution. Finally, to evaluate the robustness of our model, we also test the model on a publicly available colonoscopy video dataset and compare our results with state-of-the-art results in the literature.

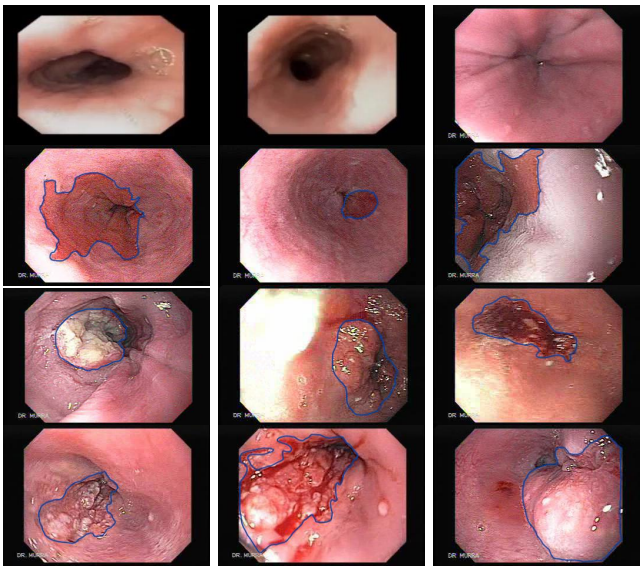


Fig. 5: Example of frames from the video dataset used in the evaluation of the proposed model. The first row shows samples from normal video frames. The second row illustrates samples from precancerous BE videos. Finally, third & forth represents cancerous samples from EAC and SCC videos respectively. The annotation by the expert is shown in blue for both the BE and EAC frames.

A. Implementation Setup and Evaluation Measures

The model is implemented using Keras Libary (Tensorflow backend) on a desktop with Intel Core i7 (3.6GHz processor) and an NVIDIA GeForce GTX1080 Ti with 11GB on a single GPU memory. The weights are initialized randomly with a gaussian distribution ($\mu = 0$, $\sigma = 0.01$). The initial learning rate was set to ($1e-5$) and drops by the factor 0.1 every 1000 iteration and used a weight decay of 0.0004.

To select the parameters of the 3D Sequential Dense-ConvLstm, different values for the *Seq-DB*= 3, 4 and 5 and growth rates (G)= 16, 24 and 32 were evaluated on the dataset as will be shown in the next section. The optimal 3D Sequential Dense-ConvLstm network performance in our model is composed of 5 dense blocks with $G = 24$. Moreover, the initial ConvLstm filter was set to include 10 frames to

capture spatiotemporal features. During implementation, we tried to include more frames but due to the limited GPU memory, the model could not handle more than 10 frames.

For the *FS-CRF* we tested different window frame t =5, 10, 15, 20 and 25 to find the best performance. As will be discussed, selecting $t = 15$ gave the best results balancing between precision and recall values. Furthermore, we set the IoU=0.7 between the two nearest frames L_x and L_y. A higher IoU value is chosen to guarantee that the detected region is the same between these two frames. The hyperparameters of the fully-connected CRF were defined in a configuration experiment using a random search on the validation data: $w1 = w2 = 1$, $\delta_\alpha = 80$, $\delta_\beta = 13$ and $\delta_\gamma = 3$. The mean-field algorithm was performed for 10 iterations.

For the evaluation, we employ the standard evaluation metrics: *Recall, Precision, F-measure* and *mean Average Precision (map)* to evaluate the performance of the model in the detection of different esophageal abnormalities from the videos compared to the ground truth annotation. The IoU (i.e. eq. (3)) is used to measure the overlap ratio between the detection results and the manual segmented gold standard.

B. Evaluation FS-CRF 3D Seq. DenseConvLstm Model

Firstly, we evaluate the overall performance of our method in detecting different abnormalities from the endoscopic videos. The results are summarized in Table I and visualized in Fig. 6 and Fig. 7. The detection model without the post-processing phase represents a good performance with a recall (88.4%), precision (89.6%) and F-measure (88.9%) which proves the efficiency of the proposed network in extracting relevant spatiotemporal feature from videos. After applying the FS-CRF postprocessing phase to the model, the results have been significantly improved to recall (93.7%), precision (92.7%) and F-measure (93.2%). The proposed FS-CRF attempts to locate abnormal regions missed in intra-frame series caused by any disturbance during movement or in nearby frames. The post-processing boosted the recall performance of the model by 5.3%. Additionally, it effectively removed false positives detected by the network improving the precision by 3.1%.

Moreover, the McNemar statistical test was conducted to validate the significance of the detection performance presented in Table I for the difference between the model with and without the FS-CRF. The test showed that the results were found to be significantly different at the level of 5% (p-value<0.05).

TABLE I: Detection results of the proposed 3D Sequential DenseConvLstm with and without (w/o) the suggested post-processing FS-CRF method

Methods	Recall	Precision	F-measure
With FS-CRF	**0.937**	**0.927**	**0.932**
W/O FS-CRF	0.884	0.896	0.889

Additionally, Fig. 6 provides different examples of our proposed detection model for the different types of abnormalities

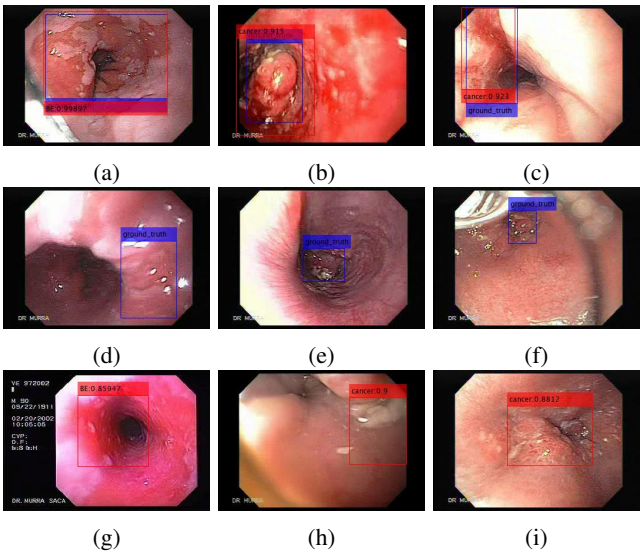


Fig. 6: Examples from the detection output of the proposed FS-CRF 3D Sequential DenseConvLstm model. The first-row illustrates samples from positive detection. The second row shows false-negative output where the model was not able to locate the abnormality. Finally, the third row represents samples from false positive detection. The three types of abnormalities: *BE, SCC*, and *EAC* are represented in First, second and third columns respectively.

(i.e. *BE, EAC, SCC*). Figs 6a through 6c represent samples of a positive detection for the three abnormal cases, showing the output results in (red bounding box) overlapping with ground truth annotation (purple bounding box). We find that our model can successfully detect the different types of abnormalities with a large IoU with the ground-truth and a high confidence score. The proposed model was not able to detect some abnormalities from different frames as shown in Figs 6d, 6e and 6f. After analyzing the missed abnormal regions, we conclude that most of the missed regions have a difficult view in the image with a relatively small abnormal area. Moreover, Fig. 6g to 6i demonstrate examples of false detection by the proposed model for BE, SCC and EAC respectively.

Furthermore, after studying the results from each video we observed that the model detected abnormalities from challenging frames (i.e as explained in section I). Fig. 7 illustrates several examples from these results. As displayed, Fig. 7a has the appearance of an examination tool and Fig. 7b has a lot of bubbles around the tumor, the model effectively detected the cancerous region properly compared to the ground truth. On the other hand, Fig. 7c and 7d has no ground truth annotation by the expert due to the blurry and fog appearance. As shown, the model successfully located these regions which confirms the robustness of the model. Extracting the spatiotemporal features from the video allowed the model to detect abnormal regions even if they appear in blurry or occluded frames.

Secondly, in Table II we report the performance of our method in locating the abnormalities for each class separately. As shown, the model had a high performance in locating cancerous regions as they have more unique properties than the

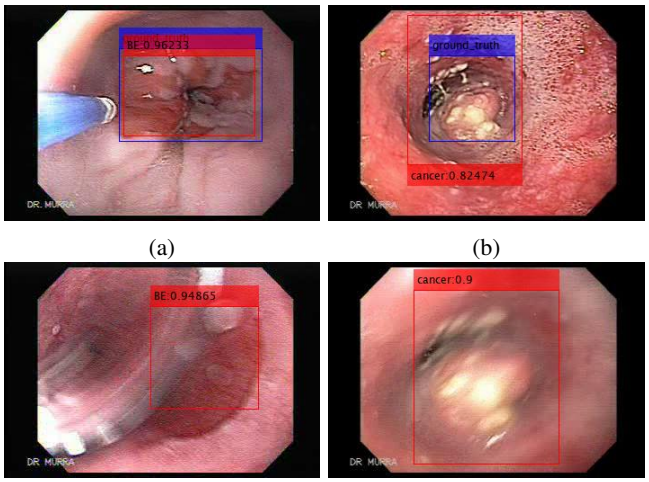


Fig. 7: Examples of detection output from challenging endoscopic frames occluding the esophageal abnormality region. (a) Tool appearance, (b) bubbles, (c) blurry, (d) fog.

precancerous stage achieving a recall of 98.2% and precision 94.9% for SCC and a recall of 94.9% and precision 91.5% for EAC. In the case of precancerous detection (BE), the model obtained a lower percentage compared to the cancerous region but still with an acceptable performance with a recall of 87.5% and a precision of 95.2%. Precancerous stages are considered more challenging to detect due to its less distinctive features and its similarity with normal regions. Moreover, the mAP for each class was measured at IoU 0.5 giving an average value of 0.7561 value which represents the efficiency of the model locating different abnormal regions compared to the ground-truth with a high confidence score.

TABLE II: Detection results for abnormality type separately using the proposed 3D Sequential DenseConvLstm with FS-CRF post-processing.

Abnormality Type	Recall	Precision	F-measure	mAP
BE	0.875	0.952	0.912	0.728
EAC	0.949	0.915	0.9324	0.758
SCC	0.982	0.919	0.949	0.782

Furthermore, to evaluate the impact of the proposed FS-CRF post-processing phase on the model, we calculate the recall and precision values after varying the window frame threshold (t). In Fig. 8 we represent the values of the precision and recall at each $t = 5$, 10, 15, 20 and 25 (i.e. t is the number the frames included before and after the selected image). Fig. 8 demonstrates that the increase of the number of frames before and after the unlabeled image improves the recall results by detecting more true positives. However, the precision value starts to decrease when including more than $t > 15$ frames. The best performance was achieved by the model at $t = 15$ (results presented in Table I).

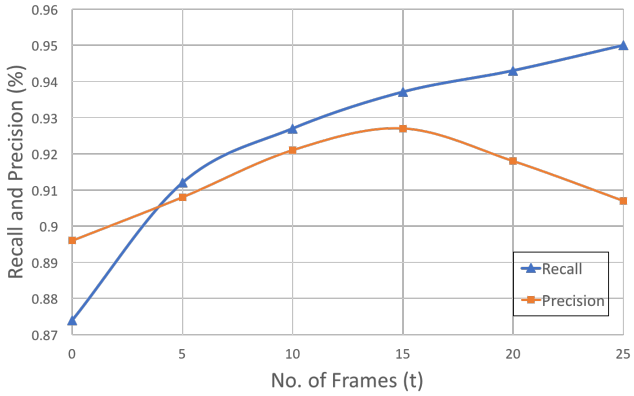


Fig. 8: The effect of changing the number of frames (t) within the window frame of FS-CRF post-processing on the precision and recall results.

C. Evaluation of 3D Sequential DenseConvLstm vs Other CNN networks

In this section, we compare the performance of the proposed network (without post-processing phase *(FS-CRF)*) with its 2D version (i.e. 2D Sequential DenseNet) which has the same architecture as the 3D Sequential DenseConvLstm but all its layers are 2D instead of 3D and replacing the ConvLstm operation with a 2D Convolutional layer. The reason for this comparison is to investigate the advantage of extracting spatiotemporal features from videos. Moreover, we compare the results of the proposed model with the results of the U-Net [50], which is a well-known network used for automatic detection/segmentation of various computer-aided diagnosis applications.

As shown in Table III, the 2D model achieved a good performance in terms of recall and precision but it still had a notable difference in the performance compared to the 3D model. The 2D obtained a comparable result in terms of the precision where the 3D model had an increase of 2.9%. On the other hand, the 3D model outperformed in detecting more abnormal regions increasing the recall value by 12.6%. This result demonstrated the efficiency of the 3D model in dealing with videos to extract spatiotemporal features that improve overall detection performance. Moreover, it can overcome the problem of challenging frames (blurry appearance, tools, bubbles, etc...) while the 2D method failed to detect them. Also, the proposed model outperformed when compared to the results of U-Net achieving the highest results for all the three metrics: *recall, precision* and *F-Measure*. Even when comparing the results of the U-Net with the 2D CNN model, it shows that our suggested network was more efficiently in extracting the relevant CNN features for better detection, producing higher F-measure value, that shows a better balance between the recall and precision results.

D. Evaluation of network configuration

Since the 3D-CNN requires a higher computational complexity than the 2D-CNN, therefore, we propose the idea

TABLE III: Performance comparison between 3D and 2D models without including the FS-CRF post-processing method and U-Net.

Methods	Recall	Precision	F-measure
Proposed Model	**0.884**	**0.896**	**0.889**
2D CNN Model	0.758	0.867	0.808
U-Net	0.731	0.859	0.789

TABLE IV: Performance of 3D Sequential DenseConvLstm and 3D Non-Sequential DenseConvLstm with different growth rate values. The number of Dense Block is fixed as 5 for both networks and growth rate G is selected from three values: 16, 24 and 32. The number of internal layers (l) is set to 5 for the 3D Non-Sequential DenseConvLstm.

Method	Params (10^7)	Acc. (%)
3D Seq. DenseConvLstm (G=16)	**8.71**	89.15
3D Seq. DenseConvLstm (G=24)	12.01	**91.10**
3D Seq. DenseConvLstm (G=32)	15.56	90.18
3D Non-Seq. DenseConvLstm (G=16)	13.71	88.23
3D Non-Seq. DenseConvLstm (G=24)	20.40	89.78
3D Non-Seq. DenseConvLstm (G=32)	27.98	90.03

of Sequential DenseNet to simplify the network architecture. The sequential structure can improve computational efficiency while preserving high performance. We compare the accuracy (Acc.) performance with the number of trained parameters (Params) for the proposed Sequential DenseNet against Non-Sequential DenseNet. The Non-Sequential DenseNet represents the architecture of the same network with the 3DCNN and Convlstm operation but with a constant number of internal layers (l) for each Dense block (i.e. similar to standard DenseNet). In Table IV, the results for both networks are presented when setting the number of dense blocks $= 5$ for both networks while varying the growth rate (G) with values: 16, 24 and 32.

As shown, even though the accuracy performance among all the 3 networks is considered comparable for accuracy results, the number of trained parameters is much reduced with the Sequential networks. Therefore, the proposed Sequential DenseConvLstm increased the network's performance with a reduced number of connections and fewer trained parameters. Additionally, the experiments showed that the 3D Seq. DenseConvLstm performs better than the non-sequential network. Increasing the number of layers at later blocks raises the weights of channels holding informative features, reduces the number of layers in earlier blocks and decreases the weights of channels with less beneficial features. The best performance among all networks was achieved for 3D Seq. DenseConvLstm at $G = 24$.

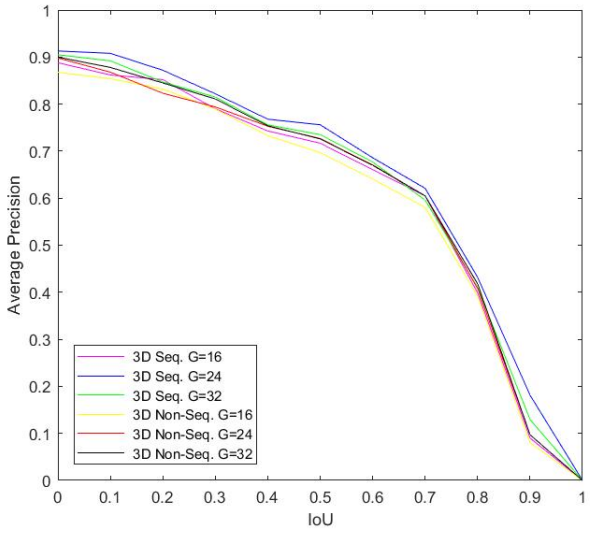


Fig. 9: AP-IoU threshold curves using different G=16, 24 & 32 values for 3D Sequential DenseConvLstm and 3D Non-Sequential DenseConvLstm networks.

Moreover, Fig. 9 represents the average precision (AP) measure as a function of the IoU threshold for the network with the different configurations. As shown, the different networks had a generally good performance in the detection of abnormal regions with the varying IoU threshold. However, the 3D Seq. DenseConvLstm at $G = 24$ was able to maintain the high performance when compared to other networks, therefore, we set our network to this configuration.

E. Comparison with other methods

To further demonstrate the effectiveness of the proposed detection model, we evaluate the model's performance on another publicly available video dataset used for the examination of the colon (*as there is no work in available in the literature for esophagus video detection to compare with*). The available dataset is known as CVC-ClinicVideoDB dataset [49], which is composed of 18 videos where each video contains a distinctive polyp appearing several times within a sequence of frames. These videos have a total number of 11954 frames with 10052 frames having polyps and annotated by experts. We compare our results with a recent method suggested in the literature by Qadir et al. [16] that uses the CVC-ClinicVideoDB dataset for evaluation. This model has two phases: First, an object detection method is used to generate region-of-interest (ROI) proposals. Secondly, a False Positive (FP) reduction unit that has a mechanism to detect FPs and correct the outliers representing the missed polyps in the sequence. The FP unit exploits the temporal dependencies between frames based on the generated regions' proposals. This model tested two object detection methods separately to generate region proposals: the Faster R-CNN [37] with the Inception ResNet [51] as the CNN backbone network and the Single Shot MultiBox Detector (SSD) [38] with the MobileNet [52] as the CNN backbone network.

Table V shows the results of our proposed model when evaluated on the CVC-ClinicVideoDB dataset where the dataset is divided randomly according to the full video into 50% training, 10% validation and 40% testing. The results are compared with the results of the model by Qadir et al. when using the Faster R-CNN and the SSD as an object detector model with one and five ROI proposals for the FP reduction unit. As shown, the results of our detection model surpassed the results by *Qadir et. al* [16] in all the evaluation measures. By using a more suitable network that can extract spatiotemporal features according to the video properties, our model increased the detection recall by 2.34% and 1.43% when compared to the other model using the Faster R-CNN with one and five ROIs respectively. For precision, our method outperformed against all the results obtained by the model [16] with a value of 96.54% which shows that our model generated much less false positives. In general, the F-measure of our model had the highest performance of 88.16% which indicates the good balance between recall and precision values.

Moreover, the proposed post-processing FS-CRF has a fast inference time in generating the updated bounding-box in each frame of the video. On the other hand, in [16] there was a delay of 0.28 seconds (sec) in displaying the detection output as the ROI of the current frame depends on the ROIs generated from the surrounding frames. In Table VI we show the Mean Processing Time (MPT) by our proposed model and compared it to the Faster R-CNN and SSD presented in [16] for processing a frame and displaying the results. As shown, the time required in our methods for each frame is an average of 0.51 *sec* which is an improved processing time compared to 0.67 *sec* for the Faster R-CNN presented in [16]. On the other hand, our method took more time compared to the proposed SSD presented by [16] which only needed 0.31 *sec*. The reason behind this is that the Faster R-CNN has an additional RPN layer to generate proposals for ROI and the network used for Faster R-CNN is much deeper than the one used for SSD. Although the SSD is much faster, it has a very low performance in terms of recall values with only 53%.

TABLE V: Comparison of the proposed model results with the method proposed by Qadir et al. [16] using the CVC-ClinicVideoDB dataset [49].

Methods	Recall (%)	Precision (%)	F-measure (%)
Proposed Model	**81.18**	**96.45**	**88.16**
Faster R-CNN (one ROI) [16]	78.84	90.51	84.27
SSD (one ROI) [16]	53.16	93.03	67.66
Faster R-CNN (five ROIs) [16]	79.75	88.50	83.9
SSD (five ROIs) [16]	53.48	92.57	67.8

VI. Conclusion

In this paper, we present a novel FS-CRF 3D Sequential DenseConvLstm model that detects different types of esophageal abnormalities (precancerous and cancerous) from endoscopic videos. The designed feature extraction network takes advantage of the nature of videos by incorporating the 3DCNN with the ConvLstm to extract spatiotemporal features (i.e. covering short and long temporal information). The proposed network proved to achieve better results when compared to the same 2DCNN network by 8.1% F-measure. Additionally, the 3D Seq. DenseConvLstm is suggested to be in a sequential matter to boost the performance of the network, reduce excessive connection and the number of trained parameters. Experiments proved that the proposed network had a fewer number of trained parameters with higher efficiency when compared to the non-sequential network. Moreover, we propose a novel post-processing phase that considers information from neighboring frames named FS-CRF to improve the overall performance. To the best of our knowledge, the presented methodology is the first to deal with the detection of esophageal abnormalities from videos instead of selected frames (still images).

TABLE VI: Comparison of the Mean Processing Time (MPT) for the proposed model with the method proposed by Qadir et al. [16] using Faster R-CNN and SSD.

	Proposed Model	**Fast R-CNN** [16]	**SSD** [16]
MTP	0.51	0.67	0.313

Future research direction includes the investigation of improving the performance of the model by detecting the small missed regions (i.e. developing a multiscale network or modifying the RPN layer), optimizing the model to enable real-time detection to be applied in the clinical routine, segmenting the detected abnormal regions. Furthermore, we will investigate methods to address different artifacts that appear in the endoscopy video (i.e. such as specularity, blood, instrument, low-contrast, illumination, and air bubbles) during the examination process.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', CA: a cancer journal for clinicians, vol. 68, no. 6, pp.394-424, 2018.

[2] R.L. Siegel, K.D.Miller, AND A. JEMAL, "Cancer statistics", CA: a cancer journal for clinicians, vol. 69, no. 1, pp.7-34, 2019.

[3] https://www.cancer.org/cancer/esophagus-cancer/detection-diagnosis-staging/survival-rates.html

[4] A. Erolu, . Can Krkolu, N. Karaolanolu, C. Tekinba, . Ymaz, and M. Baolu, "Esophageal perforation: the importance of early diagnosis and primary repair. Diseases of the Esophagus", vol. 17, no. 1, pp.91-94, 2004.

[5] N.Ghatwary, A. Ahmed, E. Grisan, H. Jalab, L. Bidaut, and X. Ye, "In-vivo Barretts esophagus digital pathology stage classification through feature enhancement of confocal laser endomicroscopy", Journal of Medical Imaging, vol. 6, no. 1, pp.014502-1-014502-12 , 2019.

[6] Flejou, J.F., 2005. Barretts oesophagus: from metaplasia to dysplasia and cancer. Gut, 54(suppl 1), pp.i6-i12

[7] A. Behrens, O. Pech, F. Graupe, A. May, D. Lorenz, C. Ell," Barrett's adenocarcinoma of the esophagus: better outcomes through new methods of diagnosis and treatment", Deutsches rzteblatt International, vol. 108, no. 18, pp. 313319, 2011 http://dx.doi.org/ 10.3238/arztebl.2011.0313.

[8] D.W. Scholvinck , K. van der Meulen, J.J. Bergman, B.L. Weusten, "Detection of lesions in dysplastic Barretts esophagus by community and expert endoscopists", Endoscopy, vol. 49, no. 02, pp. 113-20, Feb, 2017.

[9] L. A. de Souza Jr., C Palm, R. Mendel, C. Hook, A. Ebigbo, A. Probst, H. Messmann, S. Weber, J. P.Papa, "A survey on Barrett's esophagus analysis using machine learning", *Computers in Biology and Medicine*, vol. 96, pp. 203-2013 ,May, 2018. Available: https://doi.org/10.1016/j.compbiomed.2018.03.014

[10] Du, W., Rao, N., Liu, D., Jiang, H., Luo, C., Li, Z., Gan, T. and Zeng, B., 2019. Review on the Applications of Deep Learning in the Analysis of Gastrointestinal Endoscopy Images. IEEE Access, 7, pp.142053-142069.

[11] Chao, W.L., Manickavasagan, H. and Krishna, S.G., 2019. Application of artificial intelligence in the detection and differentiation of colon polyps: a technical review for physicians. Diagnostics, 9(3), p.99.

[12] Tajbakhsh, N., Gurudu, S.R. and Liang, J., 2015, April. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI) (pp. 79-83). IEEE.

[13] Misawa, M., Kudo, S.E., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., Ogawa, Y., Maeda, Y., Takeda, K., Ichimasa, K. and Nakamura, H., 2018. Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology, 154(8), pp.2027-2029.

[14] Yu, L., Chen, H., Dou, Q., Qin, J. and Heng, P.A., 2016. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE journal of biomedical and health informatics, 21(1), pp.65-75.

[15] Zhang, R., Zheng, Y., Poon, C.C., Shen, D. and Lau, J.Y., 2018. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern recognition, 83, pp.209-219.

[16] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken and Y. Shin, "Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video" in IEEE Journal of Biomedical and Health Informatics, April, 2019.

[17] S. Ji, W. Xu, M. Yang, and K. Yu, 3d convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221231, 2013.

[18] D. Maturana and S. Scherer, Voxnet: A 3D convolutional neural network for real-time object recognition, in IEEE/RSJ International Conference on Intelligent Robots and Systems.

[19] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).

[20] Zhou, Z., Chen, J., Yang, C.N. and Sun, X., 2019. Video Copy Detection Using Spatio-Temporal CNN Features. IEEE Access, 7, pp.100658-100665.

[21] Colleoni, E., Moccia, S., Du, X., De Momi, E. and Stoyanov, D., 2019. Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robotics and Automation Letters, 4(3), pp.2714-2721.

[22] Xingjian, S.H.I., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K. and Woo, W.C., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems (pp. 802-810).

[23] Nwoye, C.I., Mutter, D., Marescaux, J. and Padoy, N., 2019. Weakly supervised convolutional LSTM approach for tool tracking in laparoscopic videos. International journal of computer assisted radiology and surgery, 14(6), pp.1059-1067.

[24] Zhu, H., Vial, R. and Lu, S., 2017. Tornado: A spatio-temporal convolutional regression network for video action proposal. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5813-5821).

[25] Mathai, T.S., Gorantla, V. and Galeotti, J., 2019, October. Segmentation of Vessels in Ultra High Frequency Ultrasound Sequences Using Contextual Memory. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 173-181). Springer, Cham.

[26] Zhu, G., Zhang, L., Shen, P., Song, J., Shah, S.A.A. and Bennamoun, M., 2018. Continuous gesture segmentation and recognition using 3dcnn and convolutional lstm. IEEE Transactions on Multimedia, 21(4), pp.1011-1021.

[27] Huang, J., Li, Y., Tao, J., Lian, Z. and Yi, J., 2018, April. End-to-end continuous emotion recognition from video using 3D ConvLSTM networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6837-6841). IEEE.

[28] G. Huang, Z. Liu, L. van der Maaten and K. Weinberger,"Densely connected convolutional networks". In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708, 2017, 2016

[29] Sub-Challenge Early Barrett's cancer detection, [Online] https://endovissub-barrett.grand-challenge.org

[30] N. Ghatwary, A. Ahmed, and X. Ye. "Automated Detection of Barrett's Esophagus Using Endoscopic Images: A Survey", in *Medical Image Understanding and Analysis MIUA 2017.*, Edinburgh, UK, pp. 897-908, June 2017.

[31] Domingues, I., Sampaio, I.L., Duarte, H., Santos, J.A. and Abreu, P.H., 2019. Computer vision in esophageal cancer: a literature review. IEEE Access, 7, pp.103080-103094.

[32] R. Mendel, A. Ebigbo, A. Probst, H. Messmann, and C. Palm, "Barretts esophagus analysis using convolutional neural networks" In Bildverarbeitung fr die Medizin, pp. 80-85, 2017.

[33] S. Van Riel, F. Van Der Sommen, S. Zinger, E.J. Schoon, and P.H de With, " Automatic detection of early esophageal cancer with CNNS using transfer learning", In 2018 25th IEEE International Conference on Image Processing (ICIP), pp. 1383-1387, October, 2018.

[34] N. Ghatwary, M. Zolgharni and X. Ye, "Early esophageal adenocarcinoma detection using deep learning methods", International journal of computer assisted radiology and surgery, vol. 14, no. 4, pp.611-621, 2019.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation", IEEE transactions on pattern analysis and machine intelligence, vol.38, no. 1, pp. 142-158, 2016.

[36] R. Girshick, "Fast R-CNN", IEEE International Conference on Computer Vision (ICCV), 2015.

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pp. 9199, 2015.

[38] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pp. 2137. Springer, 2016

[39] N. Ghatwary, X. Ye, and M. Zolgharni, "Esophageal abnormality detection using DenseNet based Faster R-CNN with Gabor features". IEEE Access, vol. 7, pp.84374-84385, 2019.

[40] N. Ghatwary, M. Zolgharni and X. Ye, "GFD Faster R-CNN: Gabor Fractal DenseNet Faster R-CNN for Automatic Detection of Esophageal Abnormalities in Endoscopic Images". In International Workshop on Machine Learning in Medical Imaging Springer, Cham, pp. 89-97, October, 2019.

[41] Huang, G., Liu, S., Van der Maaten, L. and Weinberger, K.Q., 2018. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2752-2761).

[42] T. Akilan, Q.J. Wu, A. Safaei, J. Huo, and Y. Yang, "A 3D CNN-LSTM-Based Image-to-Image Foreground Segmentation", IEEE Transactions on Intelligent Transportation Systems, 2019.

[43] Karpathy, A., Johnson, J. and Fei-Fei, L., 2015. Visualizing and understanding recurrent networks. arXiv preprint arXiv:1506.02078.

[44] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2403-2412, 2018.

[45] X. Xu, F. Zhou, B. Liu, D. Fu, and X. Bai, "Efficient Multiple Organ Localization in CT Image using 3D Region Proposal Network", IEEE transactions on medical imaging, 2019.

[46] P. Krhenbhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials", In Advances in neural information processing systems, pp. 109-117, 2011.

[47] http://gastrointestinalatlas.com, Last accessed: 1 December 2018.

[48] Chollet F (2015) Keras. https://keras.io/

[49] Q. Angermann, J. Bernal, C. Snchez-Montes, M. Hammami, G. Fernndez-Esparrach, X. Dray, O. Romain, F.J. Snchez, and A. Histace, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis", In Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures, pp. 29-41, 2017.

[50] Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[51] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, vol. 4, pp. 12, 2017.

[52] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.