Knowledge-Based Systems 222 (2021) 106971

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Recurrent generative adversarial networks for unsupervised WCE video summarization

Libin Lan [a], Chunxiao Ye [a,b,*]

[a] College of Computer Science, Chongqing University, Chongqing 400044, China
[b] Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing 400044, China

ARTICLE INFO

Article history:
Received 30 November 2020
Received in revised form 14 March 2021
Accepted 17 March 2021
Available online 20 March 2021

Keywords:
Wireless capsule endoscopy
Video summarization
Variational autoencoder
Pointer network
Generative adversarial network
De-redundancy mechanism

ABSTRACT

Wireless capsule endoscopy (WCE) produces amounts of redundant images in one examination, which is very laborious and time-consuming for a physician to review these. It has been extremely needed for a technique that automatically produces a shortened and informative WCE video summary from its original video. This paper considers unsupervised WCE video summarization, and casts it as a sequence-to-sequence learning problem. Our key idea is to learn a deep summarizer network to minimize information loss between training videos and their summaries, in an unsupervised way. To this end, we propose a hybrid yet effective unsupervised WCE video summarization method using long short-term memory (LSTM), variational autoencoder (VAE), pointer network (Ptr-Net), generative adversarial network (GAN), and de-redundancy mechanism (DM) etc. techniques. The proposed model termed Adv-Ptr-Der-SUM adopts a generative adversarial framework, consisting of a summarizer and a discriminator. The summarizer is the VAE-based LSTM architecture with Ptr-Net and DM that aims to learn the conditional probability of output sequence and provide a compact summary. The discriminator is another LSTM aimed at distinguishing between the original video and reconstructed video from the summarizer. The summarizer and discriminator are adversarially trained to optimize the summarizer and produce optimal WCE video summary. Extensive experiments on our WCE-2019-Video dataset show that our model can outperform other video summarization approaches by a large margin in both supervised and unsupervised settings. Also, the proposed model is applied to two public multimedia benchmark datasets, verifying its effectiveness and generality, and demonstrating that it can achieve a competitive result.

1. Introduction

Wireless capsule endoscopy (WCE) is a new technique available for investigation of the digestive tract, particularly the whole of the small bowel where the conventional endoscopy is unable to reach [1]. WCE has proved to be an irreplaceable tool in the diagnosis and management of small bowel disorders since its approval in 2001 [2]. Before WCE is utilized, completely examining the entire small bowel is difficult. In contrast to conventional procedures (e.g., push enteroscopy, colonoscopy, and gastroscopy) to diagnose the gastrointestinal (GI) diseases [1,3], WCE has the following several advantages [1,4–9], for a complete exploration of small bowel, that (1) it allows for diagnostic and evaluation without invasive and pain; (2) it goes through the entire small intestine without restriction and easily takes endoscopic imaging of the entire small bowel; (3) the captured video images are transmitted wirelessly to a data-recording device and then used to examine off-line later to make diagnostic decisions by clinician; (4) its role has been analyzed and evaluated in many small bowel diseases such as obscure GI bleeding, Crohn's disease, GI polyposis syndromes, and small bowel tumor; (5) it has proved to be an extraordinarily safe device. Literature [10] reported in 2006 that over 340k capsules have been deployed worldwide with no reported deaths and with few side effects.

Although the WCE has made much success in noninvasive GI disease detection, there are still some challenges associated with this advanced technology. **One main problem** is that the WCE produces about 55,000 images of the whole digestive tract during an examination, which have large amounts of redundant (e.g., high similarity) or uninformative (such as intestinal juice, bubbled and undigested residue) frames, as illustratively shown in Input part of Fig. 1. This is very time-consuming and fatiguing for a physician to review WCE video sequences [11]. **Furthermore**, the collected WCE images with a variety of abnormalities

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author at: College of Computer Science, Chongqing University, Chongqing 400044, China.
E-mail address: yecx@cqu.edu.cn (C. Ye).

https://doi.org/10.1016/j.knosys.2021.106971
0950-7051/

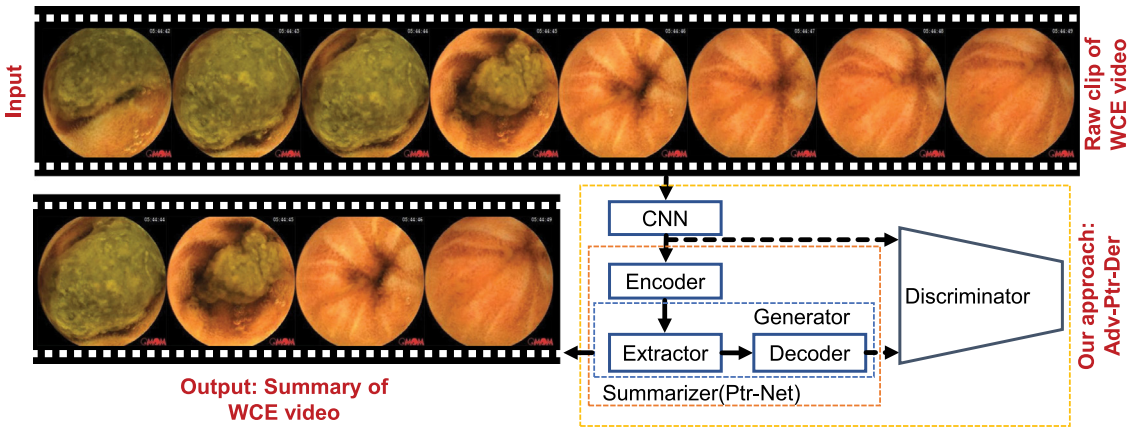


Fig. 1. Overview of the Adv-Ptr-Der-SUM model. The summary video is selected by the summarizer using pointer network from the input original video. We use a generative adversarial framework for optimizing the summarizer so as to produce optimal WCE video summary. As shown in the illustration, there are amounts of redundant or uninformative frames in WCE videos. Top row, some redundant frames with high similarity in original video; Left in bottom row, the summary produced by our Adv-Ptr-Der-SUM model. In this work, we consider using the proposed de-redundancy mechanism to eliminate redundant frames.

(that is, informative frames containing various GI disease lesions) typically account for only a small percentage of the entire WCE video images [12]. This makes it difficult to find abnormal and valuable frames quickly. Thus, it is awfully beneficial to find computer-aided diagnosis method to help clinicians identify problematic images as quickly as possible, such that it can reduce clinicians' workload and improve the efficiency. **Moreover**, since the introduction of the first capsule endoscope in clinical practice (i.e., the PillCam® developed by Given Imaging [1]) in 2001, a variety of methods have been proposed, including algorithms for detecting all types of abnormalities and reducing the reviewing time [13], as well as applications of recent artificial intelligence techniques, specifically machine learning and deep learning, are beginning to emerge in gastrointestinal endoscopy [14,15]. These can be implemented and enhance the diagnostic yield of WCE, providing the most promising results. However, in actual clinical practice, the clinician would always like to confirm the detection results generated by the computer techniques and not take any risk of missing important something in the WCE examination. All these problems motivate us to explore new computational methods that can provide a compact, shortened, and informative WCE video summary to clinician so as to reduce the time spent in the examination. We believe that it is critical to explore artificial intelligence techniques to processing and analysis of endoscopic video images.

Depending on the extensive and promising application, many computer-aided methods have been proposed to classify and detect various types of WCE diseases. That is, these mainly focus on identification of one specific abnormality or multiple abnormalities on WCE video images, e.g., bleeding [8], polyps [7], tumors [16], and ulcers [17] etc., only for one specific abnormality, and [6,18–20] for multiple abnormalities. For more details on computer-aided methods applied to WCE diseases recognition/detection and analysis, one can refer to [21,22]. However, these methods cannot provide the gist of the entire WCE video to clinician, so that important something may be missed in one examination. To deal with these problems, other approaches aimed at automatic summarization of medical videos, have also been proposed to get a hint on the entire semantic content, so as to save time of clinician inspection, or enable video sequence analysis. These existing methods involve non-negative matrix factorization-based unsupervised WCE video summarization [23–25], WCE video summarization using the factorization analysis based on sliding window singular value decomposition [26], visualization of multiple consecutive frames and highlight extraction, i.e., Quadview and Quickview modes of Rapid® 5 Access software, respectively [27,28], adaptive control [29], epitomized summarization integrating local context preservation and expectation maximization (EM) learning methods [30], feature extraction and image registration techniques-based panoramic visual summaries [31], deep learning-based noisy content removal and organ segmentation [32], summarizing CE videos into multiple classes [33], uniform sampling, motion analysis [34–39], similarity [40–42], filter [43], and clustering [44–46]. However, these methods have their limitations, e.g., for Quadview method, it is very difficult for clinician to fully perceive the content of more than four different images simultaneously. Therefore, an automatic summarization method for analyzing and understanding WCE videos are essential so as to allow for a quick filtering of undesired content and an easier browsing of the gist.

As Schoeffmann et al. point out, live endoscopic videos contain much highly similar content, exhibiting minimal obvious shot boundaries, and are also corrupted by unpredictable interruptions [47]. Thus, common shot-detection and key frame extraction methods like those [48] proposed by Smeaton et al. for traditional multimedia cannot be used with endoscopic video. That needs a new key frame extraction approach for this special medical video domain which we call it medical multimedia video. In addition to that, as we all know, video summarization is also a challenging problem in the traditional multimedia user video field and related technologies have gained increasing attention, leading to various methods proposed to help efficiently browse, manage and retrieve video contents and facilitate large-scale video distilling [49–55]. These are typically based on learning techniques to summarize video, including unsupervised and supervised methods. And their performance was usually evaluated on user videos [56,57]. That is, the study subjects of these methods are user videos rather than medical videos or images. Thus, in this work, we attempt to propose an effective resolution for our endoscopic video summarization problem via using these techniques applied to the traditional multimedia user video.

In this paper, we focus on WCE video summarization, which is defined as raw unedited video data from patient. Such data often contains some predictable abnormal patterns (that is, informative frames), e.g., some abnormal frames including bleeding, tumors, and ulcers, but is raw and therefore often long and redundant. In order to eliminate redundant WCE frames and provide a skim for clinician, we propose a video summarization algorithm for WCE video summaries. We take into account unsupervised WCE video summarization and cast it as a sequence-to-sequence (seq2seq) learning problem. Given a sequence of WCE video frames, our

goal is to acquire a short representative synopsis, which best summarizes the original WCE video.

Our work is inspired by the recent success of applying long short-term memory (LSTM) [58], variational autoencoder (VAE) [59], attention mechanism (AM) [60,61] and generative adversarial networks (GAN) [62] to user-oriented video summarization [49,51–53,63–71], as well as coverage mechanism [72] and pointer network (Ptr-Net) [73] to structured prediction problems such as text summarization [74–76] and machine translation [77]. Recently, a similar idea of unsupervised encoder–decoder [78] is also applied to video classification and video captioning. Also, a cross-layer attention mechanism [79] is used to model the multi-level information from different convolutional layers, which benefits the task of action recognition. While our key idea is to learn a deep summarizer network with de-redundancy mechanism (DM) to minimize information loss between training videos and their summarizations, in an unsupervised way. The idea of de-redundancy mechanism is derived from coverage mechanism. Our approach consists of a summarizer (pointer network) and a discriminator, which use recurrent neural networks (RNNs) especially LSTMs that excel at modeling long-range structural dependencies. We name such an adversarial learning with pointer network and de-redundancy mechanism for WCE video summarization as Adv-Ptr-Der-SUM.

An illustration of the proposed framework is given in Fig. 1. The summarizer aims to select key frames. The generator and discriminator adversarially learn to enforce both the information completeness and compactness of summaries. This ensures that the summaries capture enough key video representation from a global perspective rather than a trivial randomly shorten sequence. Existing approaches to generating summaries of WCE video do not take advantage of deep learning technique, or only use it to feature extraction, e.g., [45]. To the best of our knowledge, our work is the first to apply long short-term memory (LSTM), variational autoencoder (VAE), pointer network (Ptr-Net), generative adversarial networks (GAN), and de-redundancy mechanism (DM) to WCE video summarization.

Quantitative and qualitative evaluations on our WCE-2019-Video dataset, as well as two public benchmarks: SumMe [56] and TVSum [57] demonstrate the effectiveness of our proposed method.

In summary, the main contributions in this paper are given as follows:

1. We first explore and propose a new approach to unsupervised WCE video summarization by combining variational autoencoder (VAE), pointer network (Ptr-Net), and generative adversarial networks (GAN) techniques, which are typically used to user multimedia videos. These techniques, especially Ptr-Net, are first applied to WCE video summarization.
2. We first use de-redundancy mechanism (DM) to eliminate redundant frames in both user videos and medical videos, which is different from the previous proposed various diversity models. The idea is inspired by coverage mechanism applied to solving the repetition problem in text summarization.
3. We present a new dataset, called WCE-2019-Video, that allows for a repeatable evaluation of WCE video summarization methods. To our knowledge, it is the first dataset that can be used for future WCE related researches.

The remainder of this paper is organized as follows. Section 2 reviews prior work. Section 3 specifies our approach and training process for WCE video summarization. Section 4 briefly introduces our dataset and gives experimental details and results, and finally Section 5 presents our conclusion.

2. Related work

In the section, we will review the related from the following four aspects. (1) problem formulations, (2) traditional approaches and deep architectures to endoscopic video summarization, (3) supervised vs. unsupervised techniques for endoscopic video summarization, (l4) approaches to WCE redundant frames elimination.

Video Summarization Formulations. In this paper, we will broadly divide video summarization into user multimedia and medical multimedia video summarization. Currently, some video summarization techniques, particularly deep learning techniques, are typically applied to user multimedia video. In this work, we explore similar techniques to endoscopic video summarization. In the multimedia video domain, given an input video, video summarization aims to produce a shortened version that highlights the representative video frames. Various prior work has proposed solutions to this problem, including several classical formulations: subset selection [52,64,80], structured prediction [52,53], sequential decision making [50], and seq2seq learning [49,66,81]. According to the dividing method of problem formulations, our work is most closely related to seq2seq learning problem, selecting some important pieces of information in the original video to form a compact summary by pointer network. Here, we aim to select key frames for summarizing a WCE video. To our best of knowledge, there is no related work applying pointer network to WCE video summarization so far.

Traditional Approaches and Deep Architectures to Endoscopic Video Summarization. Similarly to the traditional approaches to multimedia video, some generic methods applied to medical video also mainly focus on four main procedures to creating a video summary: shot segmentation [82–84], key frame extraction [44–46,82,83,85–89], similarity measure [45,46,90] and the generation of summaries. However, except for [45] which uses Siamese neural network to feature extraction and similarity measure, existing approaches to generating summary do not take advantage of deep learning. Moreover, as pointed out by Schoeffmann et al. [47], common shot detection and key frame extraction methods cannot be used for these endoscopic videos containing unedited and highly similar content. Thus, we resort to a deep architecture that can effectively capture the short- and long-range dependencies among sequential frames in our WCE video, so as to derive both representative and compact medical video summaries.

Supervised vs. Unsupervised Medical Video Summarization. From the learning perspective, video summarization methods can be roughly divided into supervised and unsupervised approaches. Supervised methods use human annotations of key frames during training to optimize their models so as to minimize loss with respect to this ground truth, and finally determine which frames or shots are more important to be selected. However, for some practical applications in a certain domain, it may be impossible to provide reliable and sufficient human annotations (e.g., w.r.t. endoscopic video summarization, to the best of our knowledge, as of yet, there is no publicly available benchmark). These domains have been addressed with unsupervised methods usually using manually defined heuristic criteria to extract key frames or key shots [91]. In this work, we explore VAE and GAN to unsupervised WCE video summarization because their properties are well-suited to our application, such as scalability, easy parallelization, and seamless integration with LSTM cells. To the best of our knowledge, only a few works adopted VAE and GAN to medical video summarization so far. With respect to deep learning and GAN in medical image analysis, readers can refer to [22,92–98].

Approaches to WCE Redundant Frames Elimination. Some early work has been devoted to WCE redundant frames elimination and video summarization [23–25,30,37,44,45,85,86,90,99–

Table 1
Some abbreviations with the corresponding full names (a). The notations used in our method and corresponding descriptions (b).

(a) List of abbreviations

Abbreviation	Full name	Abbreviation	Full name
WCE	Wireless capsule endoscopy	RNN	Recurrent neural network
Ptr-Net	Pointer network	VAE	Variational autoencoder
GAN	Generative adversarial network	DM	De-redundancy mechanism
LSTM	Long short-term memory	AM	Attention mechanism
seq2seq	Sequence-to-sequence	GI	Gastrointestinal

(b) List of notations

Symbol	Description	Symbol	Description
$\mathcal{V}$	Original video	$\mathcal{L}$	Loss function
v	Video frame, a vector	ω	DM parameters
x	Feature of $\mathcal{V}$, a vector	$\mathbb{E}$	Expectation
n	The total number of frames, a scalar	h	LSTM hidden state
$p(s)$	A probability distribution over a continuous variable s, e.g., normal distribution	ϕ	Variational parameters
$\hat{x}$	Feature of reconstruction frames, a vector	∇	Using Stochastic Gradient to update parameters
s	A latent representation of feature of frames, a latent vector	$D_{KL}(\cdot)$	KL divergence
$q(s\|x)$	The probability distribution of observing s given x, as a probabilistic encoder	θ	Generative parameters
$p(x\|s)$	The conditional generative distribution for x given s, as a probabilistic decoder	φ	GAN parameters

101] or other medical video summarizations [47,82,83,87,102, 103] . But these methods almost not adopt deep learning techniques and attention mechanism to eliminate large of redundant frames. Inspired by the recent success of attention mechanism in video summarization [49,65,66,91] and coverage mechanism [72, 75] for solving the repetition problem of generated text summarization, as well as driven by some much-needed methods to redundancy reduction for WCE or other medical videos [37,104, 105], a de-redundancy mechanism (DM) is proposed to reduce the vast redundancy produced in one examination to generate a compact and informative summary. This de-redundancy mechanism is in fact an auxiliary attention mechanism to keep track of the attention history and help attention model adjust future attention. Its thought derives from the coverage model of [72].

3. Methods

In this section, we start by giving a brief overview of the existent key techniques employed by our approach. We then describe the main building blocks of this work in the following Sections 3.1–3.6. Here we also list some abbreviations with the corresponding full names across the full paper, as well as a concise reference describing the notation used throughout our method in Table 1.

Pointer Network (Ptr-Net) [73] is one kind of special attention-based seq2seq neural architecture, in which the decoder uses soft attention mechanism as a pointer to select a member from the input sequence as the output. By learning the conditional probability of an output sequence over the input, Ptr-Net solves the problem of variable size output sequence, of which the size is equal to the length of the input. A typical seq2seq model consists of two RNNs: an encoder and a decoder. In Ptr-Net, an encoding LSTM converts the input sequence to a latent representation that is fed to the generating network (decoder). For notation purposes, given input sequences x_i and output sequences $\hat{x}_j$, let the encoder and decoder hidden states be h_i^e and h_j^d, and we compute the attention vector at each output time j as follows:

$$\begin{aligned} m_j(i) &= w^T \tanh(Wh_i^e + W_0 h_j^d), i \in (1, \ldots, n) \\ \alpha_j(i) &= \text{softmax}(m_j(i)), i \in (1, \ldots, n) \end{aligned} \tag{1}$$

where $\alpha_j(i)$, derived from $m_j(i)$ after softmax normalization, is an output distribution over the input, telling the decoder where to look to produce the next frame, and w, W, W_0 are learnable parameters. Different from the vanilla seq2seq problem with attention mechanism, which typically uses the attention distribution $\alpha_j(i)$ to produce a weighted sum of the encoder hidden states h_i^e, from which predictions are made and which frame is feed to the next time step, the Ptr-Net explicitly uses the attention distribution $\alpha_j(i)$ as pointers to the all input frames i:

$$p(\hat{x}_j|\hat{x}_{1:j-1}, x) = \alpha_j(i). \tag{2}$$

For further details on Ptr-Net, please refer to [66] and [73]. Pointer network has been successfully applied to text summarization [74–76]. Recently there is also a few works [66] attempting to explore this thought for video summarization. But different from [66] using Ptr-Net to output tuples of the starting and ending points of selected fragments, we use it to extract important frames from original video.

Variational Autoencoder (VAE) [59] consists of two neural networks. One encodes an observed data sample to an unobserved latent variable , and one decodes the latent variable back to data space, both of which can be formalized as follows, respectively:

$$s \sim \text{Enc}(x) = q_\phi(s|x), \hat{x} \sim \text{Dec}(s) = p_\theta(x|s). \tag{3}$$

The VAE regularizes the encoder by imposing a prior over the latent distribution $p(s)$. It is typical to let the prior variable be a standard normal distribution $s \sim \mathcal{N}(0, I)$, where I is the identity matrix. Similarly, $p_\theta(x|s)$ identifies the conditional generative distribution for x. $q_\phi(s|x)$ is the approximation to the posterior of the generative model $p_\theta(s|x)$ which is true but intractable. ϕ and θ are two sets of parameters, that need to be updated during learning. More specific details of the VAE technique applied to WCE video summarization will be given in the corresponding subsection.

Generative Adversarial Network (GAN) [62] is a neural network for estimating generative models via an adversarial process, in which two models, a generator and a discriminator, are simultaneously trained. The generator Gen(s) maps a prior distribution s to data space, $\hat{x} = \text{Gen}(s)$ with $s \sim p_s(s)$, while the discriminator Dis(x) discriminates between the generated samples $\hat{x}$ and the true ones from true observations x, assigning the probability Dis(x) that x comes from the true data and probability $(1 - \text{Dis}(x))$ that $\hat{x}$ comes from $\hat{x} = \text{Gen}(s)$. The goal of GAN is to find a generator which fits the true data distribution while maximizing

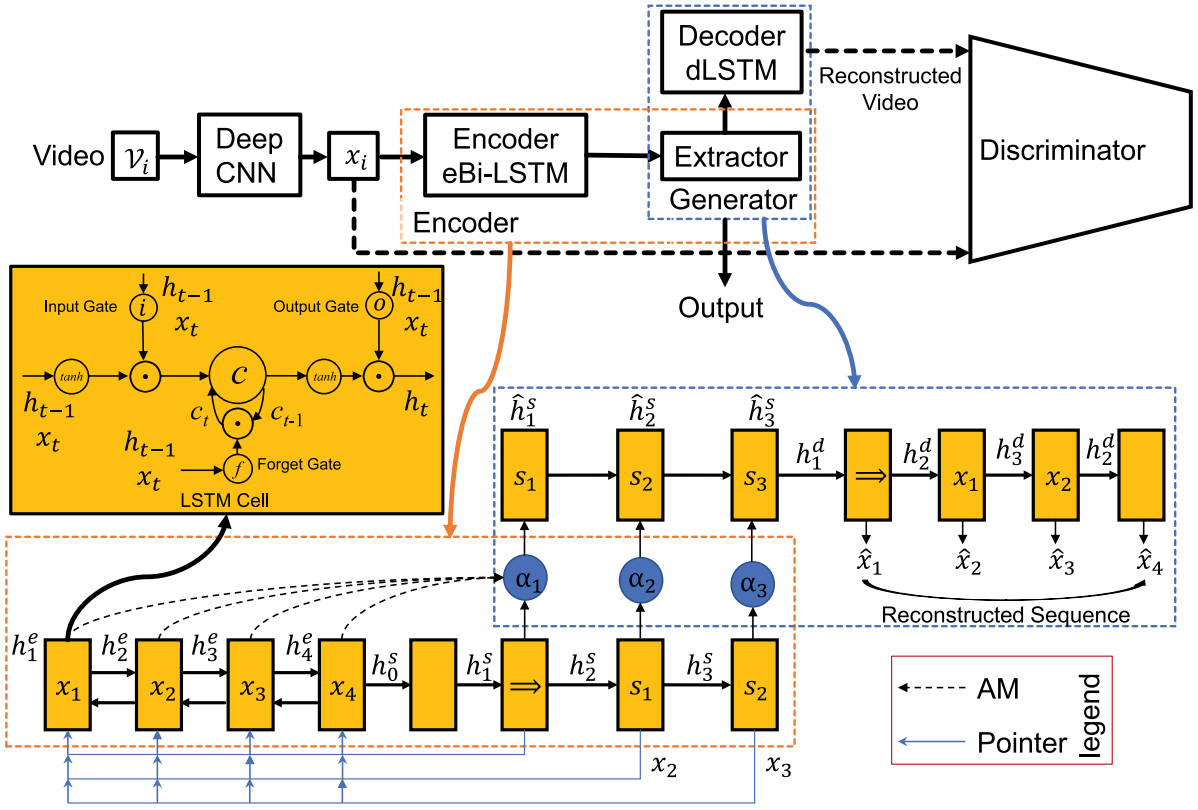


Fig. 2. Illustration of Adv-Ptr-Der-SUM architecture. For simplicity, we only depict 1-layer bidirectional LSTM as the encoder, and a 1-layer LSTM as the extractor, as well as a 1-layer LSTM with de-redundancy mechanism as the decoder. The LSTM cell is redrawn from [53], and the Ptr-Net generator is regenerated from [106]. The summarizer consists of a variational model and a generative model. AM denotes attention mechanism. Symbol ⇒ is a start symbol for each video summary or reconstructed sequence.

the probability of the discriminator making a mistake. To this end, we formulate the learning process as the following minimax optimization:

$$\min_{G}\max_{D}[\mathbb{E}_x[\log(\mathrm{Dis}(x))]+\mathbb{E}_s[\log(1-\mathrm{Dis}(\hat{x}))]], \tag{4}$$

where $\hat{x}=\mathrm{Gen}(s)$, $x\sim p_x(x)$, $s\sim p_s(s)$, $p(s)$ is a prior. We train the discriminator to maximize $\log(\mathrm{Dis}(x))$, and simultaneously train the generator to minimize $\log(1-\mathrm{Dis}(\hat{x}))$ until the generator's distribution p_s converges to the true data distribution p_x with updated parameters φ.

3.1. Architecture of our model

In this subsection, we introduce the adversarial learning with pointer network and de-redundancy mechanism (Adv-Ptr-Der-SUM) in the framework of VAE. Inspired by the extractive text/sentence summarization task in [74–76,106], we propose an Adv-Ptr-Der-SUM model that uses a 2-layer bi-directional LSTM as the encoder, a 2-layer LSTM pointer network as extractor, and 1-layer LSTM with de-redundancy mechanism as the decoder. The Adv-Ptr-Der-SUM model for WCE video summarization is a hybrid GAN model that consists of the summarizer (pointer network) and the discriminator LSTM networks, as illustrated in Fig. 2. The summarizer comprises two models: variational model and generative model, aimed at providing a variable length compact summary. The variational model (Encoder) is the variational autoencoder LSTM architecture, aimed at performing a variational inference for the true posterior distribution of summaries based on the observed video data. And the generative model (Decoder/Generator) is the conditional generative distribution over the latent vectors. The generator comprises the extractor and the decoder and reconstructs source video data. The framework of variational auto-encoder is close to the auto-encoding sentence compression model [106] and pointer generator [75].

We formulate WCE video summarization as a sequence-to-sequence learning problem. The input sequence is an original video and the output sequence is its corresponding summary, composed of the key frames. Given a video $\mathcal{V}$ of n frames, $v=\{v_i|i=1,\ldots,n\}$, each frame deep feature of the input video, $x=\{x_i|i=1,\ldots,n\}$, are extracted via a deep CNN model. The summarizer uses a bidirectional LSTM encoder (eBi-LSTM) to encode the sequence of selected frames to an extractor s, and then a decoder LSTM (dLSTM) takes s as input, and reconstructs a sequence of features, $\hat{x}=\{\hat{x}_i|i=1,\ldots,n\}$, corresponding to the input video. $p(s)$ is a prior which is typically set as the standard normal distribution $\mathcal{N}(0,I)$. The encoder model is the inference network $q_\phi(s|x)$ that takes original video x as inputs and generates extractive video frames s. The generator model is the generative network $p_\theta(x|s)$ that reconstructs x based on the extractive video frames s. Hence, the forward pass starts from the encoder to the extractor and ends at the decoder. The whole Adv-Ptr-Der-SUM model assembles a summary by selecting a subset of important frames from the original video.

We take [52] as our baseline, using a variational autoencoder (VAE) and generative adversarial networks (GANs) to perform the problem of WCE unsupervised video summarization. The key idea is that a good summary should reconstruct original video seamlessly and adopt a GAN framework to reconstruct the original video from summarized key frames. Different from [52] using a selector LSTM to output frame-level important scores, we cancel the selector, and directly adopt Ptr-Net as frame generator.

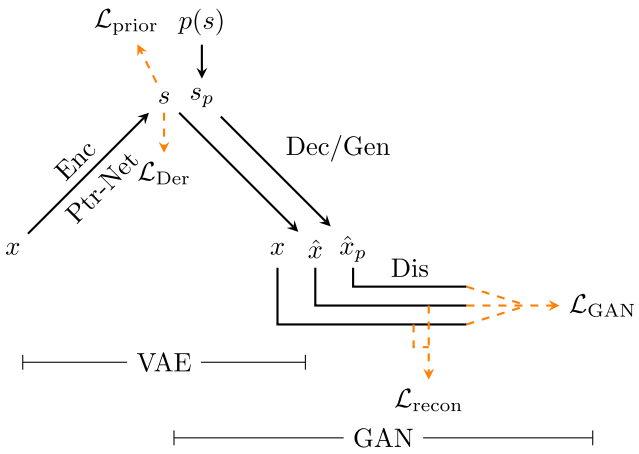


Fig. 3. All the losses used in our model. Thick solid lines denote the data flow during training; dashed lines (orange) indicate the training losses.

3.2. Variational model

Referring to [52,59,107] and [73,75,106], for the variational model (Encoder) $s \sim q_\phi(s|x)$, we use a pointer network [73] containing a bidirectional LSTM encoder that encodes the original video x to a latent variable s, and an unidirectional LSTM extractor that generates the latent video frames by attending to the encoded original video ones.

As illustrated in Fig. 2, the features of the frames x_i, in the original video are fed into the encoder, and producing a sequence of encoder hidden states h_i^e. Assuming that s_j are the features of the frames in the extractor, and h_j^s are the extractor hidden states. The variational distribution is calculated as:

$$\begin{aligned} u_j(i) &= w_3^T \tanh(W_1 h_i^e + W_2 h_j^s), \\ \alpha_j(i) &= \text{softmax}(u_j(i)), \\ q_\phi(s_j|s_{1:j-1}, x) &= \alpha_j(i), \end{aligned} \tag{5}$$

where $\phi = \{w_3, W_1, W_2\}$ are learnable parameters, h_0^s is initialized by the encoder last hidden state $h_{|x|}^e$. $\alpha_j(i)$ indicates the probability of selecting x_i as s_j, and all the frames s_j sampled from $q_\phi(s_j|s_{1:j-1}, x)$ are the subset of the frames appeared in the original video (i.e., $s_j \in x$). **For example**, $\alpha_2(1)$ indicates that the pointer selects the 2th frame x_2 in the original video as the 1th frame s_1 in the extractor, as shown in Fig. 2 (bottom).

3.3. Generative model

For the generative model (Decoder/Generator) $\hat{x} \sim p_\theta(x|s)$, $p_\theta(x|s)$ is the conditional generative distribution over the latent frames generated by the extractor. Assuming that $\hat{x}_i$ are the frames in the reconstructed video, h_i^d are the decoder hidden states, and $\hat{h}_j^s$ are the extractor hidden states. The soft attention model can be defined as:

$$\begin{aligned} \beta_i(j) &= w_6^T \tanh(W_4 \hat{h}_j^s + W_5 h_i^d), \\ \gamma_i(j) &= \text{softmax}(\beta_i(j)), \\ a_i &= \sum_j^{|s|} \gamma_i(j) \hat{h}_j^s (\beta_i(j)). \end{aligned} \tag{6}$$

The generative distribution over reconstructed video is then calculated as:

$$p_\theta(\hat{x}_i|\hat{x}_{1:i-1}, s) = \text{softmax}(W_7 a_i), \tag{7}$$

where $\theta = \{w_6, W_4, W_5, W_7\}$ are learnable parameters, h_0^d is initialized by the extractor last hidden state $\hat{h}_{|s|}^s$. Note that here the hidden state outputs $\hat{h}_j^s$ in this model are different from h_j^s in the variational model, since the information from encoder hidden states h_i^e is not involved. Thus, the parameters ϕ in the variational model are not updated by the gradients from the generative model.

3.4. Discriminator

The discriminator LSTM network learns to distinguish between both true samples and generated ones, which is trained with the generator in an adversarial learning manner. Following baseline [52], we implement the discriminator using an energy-based encoder–decoder [108] to minimize the representation error between the original video and video summarization. Similarly to the GAN presented in [62], we have that Ptr-Net generator and discriminator form the GAN framework. The GAN framework is adversarially trained so as to maximally confuse the discriminator when trying to distinguish the reconstructed videos from the original ones. Practically, in this sense, the discriminator can be viewed as a classifier estimating a distance between x and $\hat{x}$, and assigns a binary class labels to x (True) and $\hat{x}$ (False).

3.5. De-redundancy mechanism

The redundancy is a distinct characteristic for video summarization, especially WCE video summarization, for which eliminating the redundant frames is much-need. The de-redundancy mechanism is enlightened by the recent success of coverage mechanism [72], which is applied to addressing the over-translation and under-translation problem in neural machine translation (NMT) [72], and solving the repetition problem in text summarization [75], and thus we integrate it to WCE video summarization. Our proposed de-redundancy mechanism mainly aims to reduce the redundancy. To this end, we implement a de-redundancy model for all inputs i at each decoder timestep j by simply summing the attention distributions over all previous decoder timesteps:

$$r_j(i) = \sum_{j=0}^{j-1} \alpha_{j-1}(i). \tag{8}$$

Apparently, $r_j(i)$ is a distribution representing the likelihoods of the original video frames x_i being selected from the attention distributions. Note that $r_0(i)$ is initialized by a zero vector, which is because on the first timestep, none of the original video frames has been selected. The de-redundancy vector $r_j(i)$ is fed into the

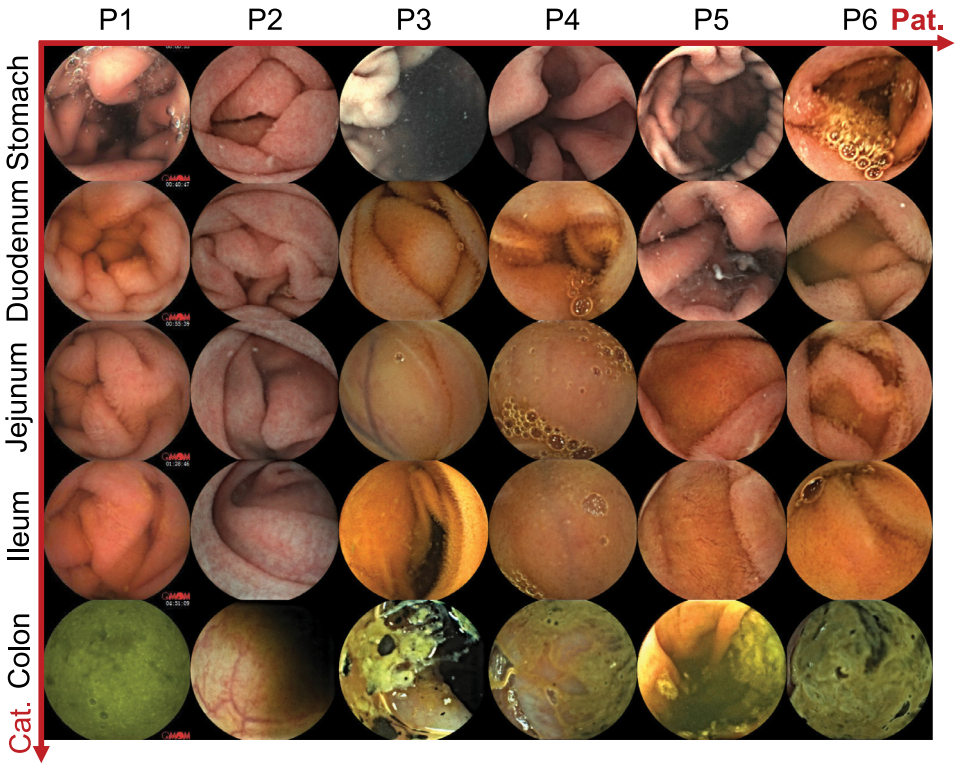


Fig. 4. WCE-2019-Video dataset contains 30 videos using 5 categories (i.e., corresponding organ of digestive tract) including: Stomach, Duodenum, Jejunum, Ileum and Colon. For the sake of using WCE-2019-Video dataset easily, based on the main component organs of digestive tract, approximate passage times of WCE through corresponding organ [109], and physicians' advice and guideline, we divided WCE-2019-Video into five parts as categories using name of corresponding organ. Each part contains 6 patients' corresponding video frames. As esophageal passage usually takes only seconds or a few minutes, this dataset does not involve the WCE images of esophagus. Furthermore, since mean small bowel transit times is approximately 235–280 min (about 4–5 h), taking about 28,800 frames, this dataset divides the whole small intestine into three parts: Duodenum, Jejunum, and Ileum so as to analyze and summarize WCE videos. Cat. axis represents category. Pat. axis denotes patient ID.

attention model (first equation) in composite Eq. (5), to help adjust next attention, expressed as:

$$u_j(i) = w_3^T \tanh(W_1 h_i^e + W_2 h_j^s + \omega r_j(i)), \tag{9}$$

where ω is a learnable parameter. This ensures that the whole summarization system can avoid repeatedly attending to the same locations or a small neighbor centering at the ith original frame, since the current decision of attention mechanism $u_j(i)$ can be made by taking into account the past selected information embedded in $u_{j-1}(i)$, and thus avoid generating redundant frames.

To intuitively illustrate how the de-redundancy mechanism eliminates the redundancy, we take $x = \{x_1, x_2, x_3, x_4\}$ as an example of input video frames and define a rule that consecutive frames are more likely to be similar with each other. When the rule is applied, we may produce a de-redundancy vector $r = \{1, 0, 0, 1\}$ or $r = \{0, 1, 0, 1\}$. This means that frames x_1 and x_4 or x_2 and x_4 are selected and others are omitted so as to achieve the purpose of deleting redundant frames. Other rules such as Determinantal Point Process (DPP) [53,80] for diversity regularization can also be used.

As can be seen from the above procedure, our de-redundancy mechanism used to eliminate the redundant frames is different from other various diversity models also aimed at mitigating or reducing the redundancy in their summaries [53,80,110].

3.6. Training loss

As illustrated in Fig. 3, we design the following loss functions to train the Adv-Ptr-Der-SUM model: (1) generative loss $\mathcal{L}_{\text{Gen}}$, consisting of a prior loss $\mathcal{L}_{\text{prior}}$ and a reconstruction loss $\mathcal{L}_{\text{recon}}$, both of which are used to train the VAE-based generator; (2) adversarial loss $\mathcal{L}_{\text{GAN}}$; (3) de-redundancy loss $\mathcal{L}_{\text{Der}}$.

Generative Loss $\mathcal{L}_{\text{Gen}}$. $\mathcal{L}_{\text{Gen}}$ contains a prior loss $\mathcal{L}_{\text{prior}}$ and a reconstruction loss $\mathcal{L}_{\text{recon}}$. As mentioned above, we use VAE-based LSTM as the generator. The loss of the VAE-based generator is written as:

$$\mathcal{L}_{\text{Gen}} = D_{KL}(q(s|x) \parallel p(s)) - \mathbb{E}[\log(p(x|s))], \tag{10}$$

where the first right hand side (RHS) term is the Kullback–Leibler divergence for the prior loss,

$$\mathcal{L}_{\text{prior}} = D_{KL}(q(s|x) \parallel p(s)). \tag{11}$$

The second RHS term is the reconstruction error for the reconstruction loss $\mathcal{L}_{\text{recon}}$, which typically uses the Euclidean distance, $\|x - \hat{x}\|_2$, between input and reconstructed output. However, the recent findings [107] reveal the limitations of the simple element-wise metric. So, this work [107] presents jointly training the VAE and the GAN so as to use the hidden representations in the GAN discriminator for measuring sample similarity. We also use the same idea to measure the video distance. To this end, we let $\text{Dis}_l(x)$ denotes the hidden representation of the last hidden layer of the discriminator, corresponding to the input of the VAE, x. Therefore, we can replace the VAE reconstruction error term with a following reconstruction error expressed in the GAN discriminator:

$$\mathcal{L}_{\text{recon}} = -\mathbb{E}[\log(p(\text{Dis}_l(x)|s))], \tag{12}$$

where expectation $\mathbb{E}$ is approximated as the empirical mean of training examples. Following [52], in this paper, we consider $p(\text{Dis}_l(x)|s) \propto \exp(-\|\text{Dis}_l(x) - \text{Dis}_l(\hat{x})\|^2)$. Variational parameters

ϕ and generative parameters θ will be updated during learning. For efficient learning that can differentiate and optimize the variational lower bound (i.e., $\mathcal{L}_{\text{Gen}}$), the reparameterization trick in [59] of the $\mathcal{L}_{\text{Gen}}$ is used for stochastic gradient descent.

GAN Loss $\mathcal{L}_{\text{GAN}}$. As mentioned above, the GAN objective is to train the discriminator such that it can maximize the probability making a mistake and simultaneously encourage the generator to fit the true data distribution. Considering that using samples from $s\sim\text{Enc}(x)=q(s|x)$ may yield better results, we add a reconstructed signal to $\mathcal{L}_{\text{GAN}}$. According to Eq. (4), the final GAN loss $\mathcal{L}_{\text{GAN}}$ is as:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}=&\mathbb{E}[\log(\text{Dis}(x))]+\mathbb{E}[\log(1-\text{Dis}(\hat{x}_p))]\\&+\mathbb{E}[\log(1-\text{Dis}(\hat{x}))],\end{aligned}\tag{13}$$

where $\hat{x}_p=\text{Gen}(s_p)$ with $s_p\sim p(s_p)$, $\hat{x}\sim\text{Dec}(s)=p(x|s)$, and $x\sim p(x)$. Following [52], we use an energy-based GAN [108] to minimize the representation error between the original video and video summarization.

De-redundancy Loss $\mathcal{L}_{\text{Der}}$. Following [75], we define a de-redundancy loss $\mathcal{L}_{\text{Der}}$ to penalize repeatedly attending to the same locations on each timestep j:

$$\mathcal{L}_{\text{Der}}=\sum_{j=0}^{j}\min(\alpha_j(i),r_j(i)).\tag{14}$$

For video summarization, since it should not require uniform distribution, we only penalize the overlap between each attention distribution and the de-redundancy vector $r_j(i)$ so as to prevent repeated attention.

Overall Loss $\mathcal{L}$. We train the Adv-Ptr-Der-SUM model by using the above three losses as the overall loss function:

$$\mathcal{L}=\mathcal{L}_{\text{Gen}}+\mathcal{L}_{\text{GAN}}+\mathcal{L}_{\text{Der}}.\tag{15}$$

Note that since both decoder and generator map from s to x, θ are the shared parameters between the two.

Training Adv-Ptr-Der-SUM. Similarly to the training procedure of [52,107], we iteratively optimizes the above triple objectives. We train our hybrid model using the overall loss function $\mathcal{L}$ in Eq. (15). This is possible because not all network parameters are updated with respect to the final loss $\mathcal{L}$. Additionally, we did not consider the relative importance of each loss, since we hope that each loss can provide an equally important backpropagation signal. Thus, any hyper-parameter, balancing gradient contributions to be updated parameters, did not be adopted in the above triple criterion and final loss $\mathcal{L}$. One can refer to Fig. 3 and Algorithm 1 for the training procedure of our Adv-Ptr-Der-SUM model for all parameters to be updated.

Given the above training losses $\mathcal{L}_{\text{GAN}}$, $\mathcal{L}_{\text{Gen}}$, and $\mathcal{L}_{\text{Der}}$, we update the parameters $\{\phi,\theta,\varphi,\omega\}$ in training using the Stochastic Gradient Variational Bayes estimation [59]. Both VAE-based generator and GAN are jointly trained to maximally confuse the discriminator.

Algorithm 1: Training Adv-Ptr-Der-SUM model.

Input: Deep features of training video x
Output: Learned parameters for the Encoder ϕ, the Decoder/Generator θ, the Discriminator φ, and the De-redundancy Mechanism ω

Initialize all parameters $\{\phi,\theta,\varphi,\omega\}$
repeat
for max number of iterations **do**
$x\leftarrow$ mini-batch frame features from CNN
$s\leftarrow$ eLSTM(x) % encoding, Ptr-Net selects frames
$\mathcal{L}_{\text{Der}}\leftarrow\sum\min(\alpha,r)$
$\mathcal{L}_{\text{prior}}\leftarrow D_{KL}(q(s|x)||p(s))$
$\hat{x}\leftarrow$ dLSTM(s) % decoding, reconstruction
$\mathcal{L}_{\text{recon}}\leftarrow-\mathbb{E}[\log(p(\text{Dis}_i(x)|s))]$
$s_p\leftarrow$ samples from a prior normal distribution $s_p\sim\mathcal{N}(0,I)$
$\hat{x}_p\leftarrow$ dLSTM(s_p) % reconstruction
$\mathcal{L}_{\text{GAN}}\leftarrow\mathbb{E}[\log(\text{Dis}(x))]+\mathbb{E}[\log(1-\text{Dis}(\hat{x}_p))]+\mathbb{E}[\log(1-\text{Dis}(\hat{x}))]$
% Updates Parameters using Stochastic Gradient
$\{\omega,\phi\}\overset{+}{\leftarrow}-\nabla(\mathcal{L}_{\text{Der}}+\mathcal{L}_{\text{prior}}+\mathcal{L}_{\text{recon}})$
$\{\theta\}\overset{+}{\leftarrow}-\nabla(\mathcal{L}_{\text{recon}}+\mathcal{L}_{\text{GAN}})$ % θ are shared parameters between Dec and GAN
$\{\varphi\}\overset{+}{\leftarrow}-\nabla(\mathcal{L}_{\text{GAN}})$
end for
until convergence of parameters, $\{\phi,\theta,\varphi,\omega\}$
return $\phi,\theta,\varphi,\omega$

4. Experiments

4.1. Datasets

WCE-2019-Video Dataset. Since there is yet no public benchmark dataset for our task, we collect a new WCE video summarization dataset from the raw WCE videos supported by Chongqing Jinshan Science & Technology (Group) Co., Ltd. We name the dataset as WCE-2019-Video. Our built WCE-2019-Video dataset contains 5 categories and 30 videos (6 per category from 6 patients) collected at the first phase of the task, and other two videos (corresponding 2 categories from the seventh patient) collected at the second phase, totaling 32 videos. Each video varies from 600 to 7500 frames, with frame-level importance scores. Fig. 4 shows thumbnails of the 30 videos and their corresponding categories; Table 2 shows descriptive statistics. To our knowledge, there are no publicly available implementations or datasets which are used to evaluate WCE video summarization so far. So, the dataset can serve as first dataset to validate WCE video summarization techniques, and will be released after the review process.

Similarly to the annotation protocol via crowdsourcing in TVSum [57], the frame-level importance scores of each video within WCE-2019-Video dataset are annotated by 6 experienced clinicians. Also, the annotation criteria refer to TVSum dataset: (1)asking each participant to watch the whole video and provide an importance score to each frame from 1 (not important) to 5 (very important); (2) avoiding chronological bias. Chronological bias is the conception that humans tend to assign higher scores to the shots that appear earlier in video, simply by virtue of their temporal precedence, regardless of their actual visual quality or representativeness. To this end, following [57],we use the same method to obtain consistent scores for visually similar frames; (3) with respect to regularizing score distributions, we use the same method as [57] to regularize score distribution. For more details of these methods, readers can refer to related literatures. A single ground-truth summary is then computed by taking an average of all the frame importance scores.

Two Public Multimedia Benchmarks: SumMe and TVSum. SumMe [56] consists of 25 videos ranging from 1.5 to 6.5 min and provide multiple user-annotated summaries (by 15–18 different users) for each video in the form of shot-level importance scores, i.e., video segments rather than keyframes. The dataset covers multiple events from both first-person and third-person camera, such as cooking and sports. Moreover, it provides a single ground-truth summary in the form of frame-level importance scores (calculated by averaging the key-fragment user summaries per frame). TVSum [57] contains 50 videos collected from YouTube

Table 2
Following [57], we give descriptive statistics of WCE-2019-Video dataset. #vid denotes the number of videos per category. #frm shows the number of frames with corresponding category. P# represents the number of selected video frames of corresponding patient ID. Percentage is the ratio of P# to the total number of video frames of each patient. The total number of video frames of P1 to P7 patients are 28, 768, 35, 190, 41, 337, 34, 841, 24, 625, 59, 329, and 17,695, respectively.

Category	Descriptive statistics								
	P1	P2	P3	P4	P5	P6	P7	#vid	#frm
Stomach	2,950	1,631	2,420	2,758	2,896	2,328	1,510	7	14,983
Duodenum	921	832	830	975	845	621	–	6	5,024
Jejunum	3,000	3,250	4,630	3,950	1,573	3,609	2,405	7	20,012
Ileum	3,500	4,147	6,885	4,655	2,024	4,753	–	6	25,964
Colon	6,000	6,136	5,940	7,021	3,766	7,500	–	6	36,363
Total	16,371	15,996	20,705	19,359	11,104	18,811	3,915	32	102,346
Percentage (%)	56.9	45.5	50.1	55.6	45.1	31.7	22.1	–	–

in 10 categories defined in the TRECVid Multimedia Event Detection (5 videos per category). The videos in this dataset vary from 1 to 5 min and are annotated by 20 users in the form of frame-level importance scores. Also, the dataset provides a single ground-truth summary (computed by averaging all users' scores). Similarly to SumMe, TVSum also captures multiple visual styles and events from both first-person and third-person camera, such as grooming an animal and making sandwich.

For fair comparison, we evaluate our approach and related approaches using the single ground-truth summaries of each video of all three datasets: WCE-2019-Video, SumMe and TVSum. The single ground-truth summary is created by following the protocol described in [53,57,111].

4.2. Evaluation metric

For fair comparison with other state of the arts, we consider F-score used in [50–53,66,111,112] as an evaluation metric. Given ground-truth summary A and generated summary B, the precision (P) and recall (R) are calculated based on the length of temporal overlap between A and B as follows:

$$\mathrm{P} = \frac{A \cap B}{|B|}, \mathrm{R} = \frac{A \cap B}{|A|}, \tag{16}$$

where $A \cap B$ denotes the length of temporal overlap between them, and $|\bullet|$ indicates the temporal duration of ground-truth summary A and generated summary B. The harmonic mean F-score is then defined as:

$$\mathrm{F} = \frac{2\mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \times 100\%. \tag{17}$$

4.3. Implementation details

Following the canonical learning settings [52] and [53], we use 80% of given dataset as training set, and the remaining 20% of it as testing set for evaluating our model. For a fair comparison with previous methods [52,53,56,57,66,67], for all the three datasets, namely, WCE-2019-Video, SumMe, and TVSum with multiple human-generated summaries, we use the similar approach [52,53,56,57,66,67] to create a single ground-truth set for evaluation. Meanwhile, we conduct all experiments on five different random splits and report the average performance.

We train our Adv-Ptr-Der-SUM model on an Nvidia TitanXp graphics card using Adam [113] optimizer whose learning rates for discriminator and others are 1e-5 and 1e-4, respectively. We implement our approach using PyTorch [114]. Following the convention [50–53,66,67,111,112], we extract 1024d deep features as the descriptor of each video frame from the output of pool 5 layer of the GoogLeNet network [115] which is pre-trained on ImageNet [116]. Also, we use a 2-layer bidirectional LSTM as the encoder, and a 2-layer LSTM as the extractor, as well as a 1-layer LSTM with de-redundancy mechanism as the decoder in generator. We use 1024 hidden units at each layer in the whole model. Similarly to [52], we initialize encoder, extractor, and generator with the parameters of a pre-trained recurrent autoencoder model trained on feature sequences from original videos, which can accelerate convergence and improve the overall accuracy.

All these settings above are used on all three datasets: WCE-2019-Video, SumMe, and TVSum.

Table 3
Performance comparison on F-scores (%) of the Adv-Ptr-Der-SUM model and its ablation variants on WCE-2019-Video dataset. GAN, pointer network, and de-redundancy mechanism can be simply on/off. The supervised variant is implemented by adding the supervision signals of BCE loss to Adv-Ptr-Der-SUM. Unsup. and Sup. denote the unsupervised variants and supervised variants, respectively.

Setting	Method	WCE-2019-Video
Unsup.	Adv-Ptr-Der-SUM$_{\text{w/o-Ptr-Der (i.e., w/-Adv)}}$	**39.2**
	Adv-Ptr-Der-SUM$_{\text{w/o-Der (i.e., w/-Ptr-Adv)}}$	42.1
	Adv-Ptr-Der-SUM$_{\text{w/o-Ptr (i.e., w/-Der-Adv)}}$	41.3
	Adv-Ptr-Der-SUM$_{\text{w/o-Adv}}$	**38.5**
	Adv-Ptr-Der-SUM	**44.6**
Sup.	Adv-Ptr-Der-SUM$_{\text{w/o-Ptr-Der-sup}}$	39.5
	Adv-Ptr-Der-SUM$_{\text{w/o-Der-sup}}$	42.5
	Adv-Ptr-Der-SUM$_{\text{w/o-Ptr-sup}}$	41.7
	Adv-Ptr-Der-SUM$_{\text{sup}}$	**45.5**

4.4. Baseline

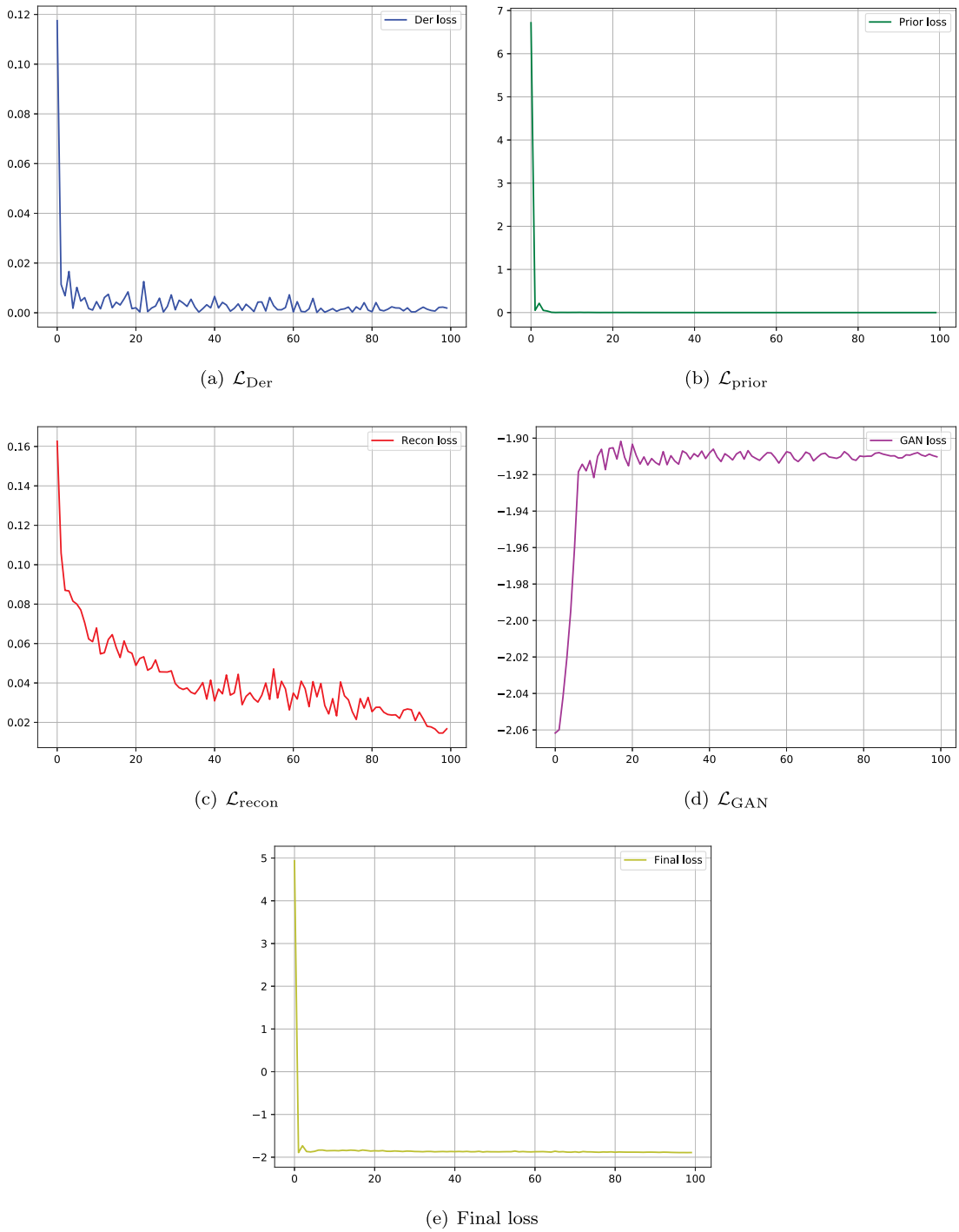
We use the VAE and GAN network structure in [52] as baseline. We choose unsupervised version of [52] to model a video. We adopt same GAN loss but drop the basic sparsity loss. Additionally, following [52], we add a binary cross entropy (BCE) loss signal between ground truth summarization positions g_j and predicted ones $q(s_j)$ for supervised learning. The objective is formalized as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{BCE}} = &-\frac{1}{|s|}\sum_{j=1}^{|s|} g_j \log(q(s_j)) \\ &+(1-g_j)\log(1-q(s_j)),\end{aligned} \tag{18}$$

where $|s|$ is the length of output sequence.

4.5. Ablation analysis

We conducted several ablation studies to analyze the contribution of each component of the Adv-Ptr-Der-SUM model. Ablation experiments are performed on our WCE-2019-Video dataset. Comparisons of ablation studies are shown by Table 3. Depending on which training loss is adopted, we consider following ablation variants of Adv-Ptr-Der-SUM.



(a) $\mathcal{L}_{\text{Der}}$ (b) $\mathcal{L}_{\text{prior}}$

(c) $\mathcal{L}_{\text{recon}}$ (d) $\mathcal{L}_{\text{GAN}}$

(e) Final loss

Fig. 5. All the loss curves of Adv-Ptr-Der-SUM model. Horizontal axis denotes training epochs. Best viewed in color with zoom-in.

Adv-Ptr-Der-SUM$_{\text{w/-Adv}}$. This variant indicates that the pointer network and de-redundancy mechanism are not adopted. It is used to verify the effects of adversarial learning to WCE video summarization. When only adversarial learning is used to trained Adv-Ptr-Der-SUM model, the model degenerates to a baseline model [52] in which we may modify some parameters.

Adv-Ptr-Der-SUM$_{\text{w/o-Der}}$. The variant denotes that the de-redundancy mechanism is not used. It drops de-redundancy loss $\mathcal{L}_{\text{Der}}$, and keeps two other training loss, which can be used to analyze the summarization performance of de-redundancy mechanism. This case is similar to ASC [106], but note that the ASC model does not use adversarial learning.

Adv-Ptr-Der-SUM$_{\text{w/o-Ptr}}$. In this case, the pointer network is not adopted. But a selector of [52] is used to select key frames, which means that training loss adopts $\mathcal{L}_{\text{recon}}+\mathcal{L}_{\text{prior}}+\mathcal{L}_{\text{sparsity}}$ in [52] instead of $\mathcal{L}_{\text{Gen}}$. Meanwhile, $\mathcal{L}_{\text{Der}}$ is kept. This variant is set to verify the summarization performance of pointer network. This case is equivalent to adding a de-redundancy signal to A-AVS [49].

Adv-Ptr-Der-SUM$_{\text{w/o-Adv}}$. This variant indicates that $\mathcal{L}_{\text{GAN}}$ is not included, which is similar to the SUM-GAN$_{\text{w/o-GAN}}$ in [52]. This case is set to show that the VAE and GAN can help to improve the performance of the model. It can be seen from Table 3 that the F-score value is the lowest in this case.

Adv-Ptr-Der-SUM. In this case, the overall loss function $\mathcal{L}$ is the overall objective for training the Adv-Ptr-Der-SUM model in

an unsupervised manner. As one can see, when all three losses are applied to training simultaneously, the highest performance can be obtained.

Adv-Ptr-Der-SUM$_{sup}$. This case is specified for the supervised setting by adding Eq. (18) between ground-truth and predicted, where ground-truth annotations of key frames are provided during training. Also, we give the other three variants of this supervised learning: Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr\text{-}Der\text{-}sup}$, Adv-Ptr-Der-SUM$_{w/o\text{-}Der\text{-}sup}$, and Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr\text{-}sup}$. These supervised variants are analogous to AALVS [66].

Comparing F-scores of Adv-Ptr-Der-SUM$_{w/o\text{-}Der}$ and Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr}$ with Adv-Ptr-Der-SUM$_{w/\text{-}Adv}$, it can be seen that the performances of the Adv-Ptr-Der-SUM model with pointer network and de-redundancy mechanism all outperform that of the model merely having adversarial learning, i.e., the baseline model in [52] by 2.9% and 2.1%, respectively. This proves that both pointer network and de-redundancy mechanism can help to improve the summarization performance.

Also, we can see that Adv-Ptr-Der-SUM$_{w/o\text{-}Der}$ outperform Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr}$ by 0.8%, which means that pointer network may be more advantageous than de-redundancy mechanism over WCE video summarization. Furthermore, Adv-Ptr-Der-SUM has the best performance in the unsupervised learning setting, exceeding Adv-Ptr-Der-SUM$_{w/\text{-}Adv}$ by 5.4%, which shows that integrating both pointer network and de-redundancy mechanism can significantly improve summarization performance.

Additionally, as one can see, Adv-Ptr-Der-SUM$_{w/\text{-}Adv}$ outperforms Adv-Ptr-Der-SUM$_{w/o\text{-}Adv}$ by 0.7%, which can prove that this model can benefit from the VAE and GAN learning.

By conducting the other four ablation studies in areas of supervised learning, we can see that our supervised variant Adv-Ptr-Der-SUM$_{sup}$, outperforms all unsupervised approaches on WCE-2019-Video dataset, as well as the other three supervised variants: Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr\text{-}Der\text{-}sup}$, Adv-Ptr-Der-SUM$_{w/o\text{-}Der\text{-}sup}$, and Adv-Ptr-Der-SUM$_{w/o\text{-}Ptr\text{-}sup}$. Adv-Ptr-Der-SUM$_{sup}$ beats Adv-Ptr-Der-SUM by 0.9%, which demonstrates that annotated data with labels and additional learning signal can improve learning. The other three experiments also verify the effects of different components of corresponding variants.

4.6. Analysis on loss curves

For the sake of space, here we report on only our training loss curves in the proposed methods to WCE video summarization in term of various relevant techniques. And simultaneously, we plot these loss curves to show that the algorithm 1 can converge toward a minimum as the network trains on the WCE-2019-Video dataset. These loss curves include five loss functions: $\mathcal{L}_{\text{Der}}$, $\mathcal{L}_{\text{prior}}$, $\mathcal{L}_{\text{recon}}$, and $\mathcal{L}_{\text{GAN}}$, as well as a final loss $\mathcal{L}$ with respect to the unsupervised method: Adv-Ptr-Der-SUM.

It can be seen from Fig. 5 that the four loss functions: $\mathcal{L}_{\text{Der}}$, $\mathcal{L}_{\text{prior}}$, $\mathcal{L}_{\text{recon}}$, and $\mathcal{L}_{\text{GAN}}$ make for a stable training, and $\mathcal{L}_{\text{GAN}}$ converges when the generator minimizes $[\log(1 - \text{Dis}(\hat{x}_p)) + \log(1 - \text{Dis}(\hat{x}))]$. Note that although the curves of $\mathcal{L}_{\text{Der}}$, $\mathcal{L}_{\text{recon}}$, and $\mathcal{L}_{\text{GAN}}$ seem to oscillate toward the minimum, the magnitudes of fluctuation are small, which are almost not more than 0.01. We believe that the fluctuation is reasonable. Also, the overall loss $\mathcal{L}$ seems to converge smoothly, which shows that our model does not suffer from severe overfitting.

4.7. Quantitative results

Comparison with other methods on WCE-2019-Video. Since our approach mainly aims to unsupervised learning, in this subsection we first compare our unsupervised variant with the other three main unsupervised methods: SUM-GAN$_{dpp}$ [52], Cycle-SUM [51] and SUM-GAN-AAE [67] on WCE-2019-Video dataset, and give a concise description for this comparison. Additionally, we also compare our supervised variant with the supervised methods: dppLSTM [53] and AALVS [66] on our dataset. These results are shown in Table 4.

Table 4
Comparison on F-scores (%) of our unsupervised and supervised variants with other unsupervised (Unsup.) and supervised (Sup.) approaches on our WCE-2019-Video dataset, respectively.

Setting	Method	WCE-2019-Video
Unsup.	SUM-GAN$_{dpp}$ [52]	38.9
	Cycle-SUM [51]	42.1
	SUM-GAN-AAE [67]	43.9
	Adv-Ptr-Der-SUM	**44.6**
Sup.	dppLSTM [53]	36.2
	AALVS [66]	44.3
	Adv-Ptr-Der-SUM$_{sup}$	**45.5**

Table 5
Comparison on F-scores (%) of our unsupervised variant with other unsupervised approaches on SumMe and TVSum. Our approach outperforms other existing methods except for CSNet and SUM-GAN-AAE which they reported 51.3% and 56.9% on SumMe, and 58.8% and 63.9% on TVSum, respectively.

Method	SumMe	TVSum
SUM-GAN$_{dpp}$ [52]	39.1	51.7
DR-DSN [112]	41.4	57.6
CSNet [50]	51.3	58.8
Cycle-SUM [51]	41.9	57.6
SUM-GAN-AAE [67]	56.9	63.9
Adv-Ptr-Der-SUM	**43.6**	**58.3**

Table 6
Comparison on F-scores (%) of our supervised variant with other supervised approaches on SumMe and TVSum. This variant performs the best on TVSum, and merely slightly lower than CSNet on SumMe.

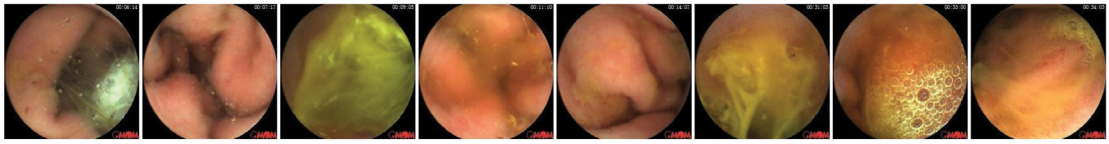
Method	SumMe	TVSum
dppLSTM [53]	38.6	54.7
SUM-GAN$_{sup}$ [52]	41.7	56.3
DR-DSN$_{sup}$ [112]	42.1	58.1
CSNet$_{sup}$ [50]	48.6	58.5
A-AVS [49]	43.9	59.4
AALVS [66]	46.2	63.6
Adv-Ptr-Der-SUM$_{sup}$	**47.7**	**64.5**

Since we could not find the authors' implementation, we re-implemented the other two main unsupervised methods: SUM-GAN$_{dpp}$ [52] and Cycle-SUM [51] in PyTorch where SUM-GAN$_{dpp}$ is reproduced based on the PyTorch Implementation of SUM-GAN.[1] Also, the code of SUM-GAN-AAE [67] is publicly available.[2] All three methods use the same settings from their published papers. The results of comparison on WCE-2019-Video between our Adv-Ptr-Der-SUM and two other methods are shown in Table 4. One can see that the summarization performance of Adv-Ptr-Der-SUM is the best with respect to other three unsupervised methods, which outperforms SUM-GAN$_{dpp}$, Cycle-SUM, and SUM-GAN-AAE by 5.7%, 2.7%, and 0.7% respectively. This demonstrates a huge advantage of unsupervised variant of our approach over existing techniques.

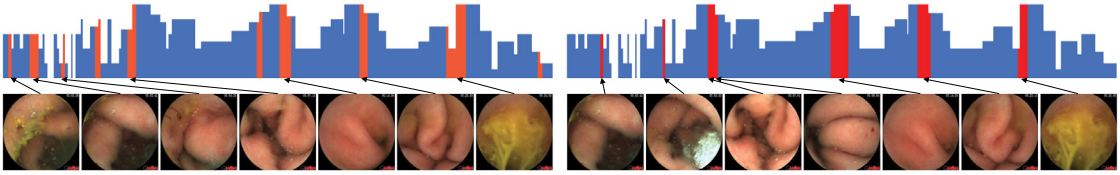
Also, we evaluate the existing supervised technique: dppLSTM [53] and AALVS [66] using WCE-2019-Video. The codes of both

[1] https://github.com/j-min/Adversarial_Video_Summary, (last accessed on Feb. 20, 2019).

[2] https://github.com/e-apostolidis/SUM-GAN-AAE, (last accessed on Jan. 19, 2020).



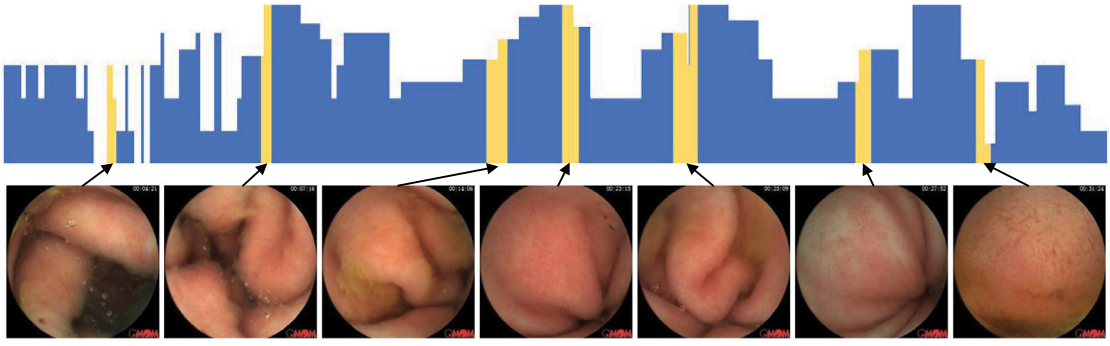
(a) Sample frames from video #31 (indexed as in WCE-2019-Video)



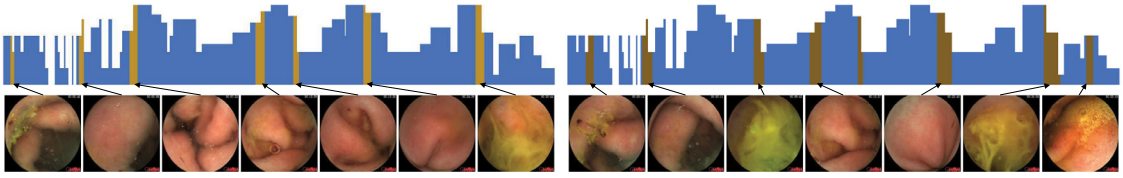
(b) Adv-Ptr-Der-SUM on video #31 of WCE-2019-Video dataset (F-score=58.5%)

(c) Adv-Ptr-Der-SUM$_{sup}$ on video #31 of WCE-2019-Video dataset (F-score=62.3%)

Fig. 6. Example summaries generated by our unsupervised and supervised variants from a sample video in WCE-2019-Video. The blue bars show the annotation importance scores. The colored segments correspond to the selected parts using different methods.

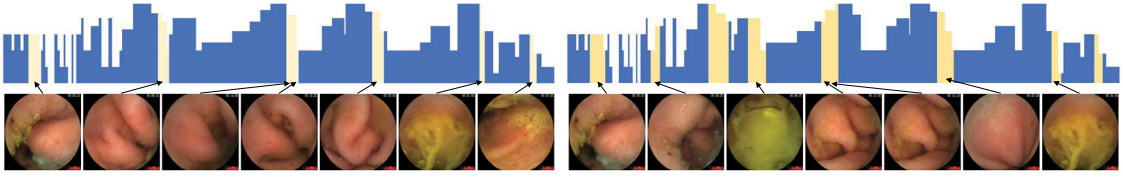


(a) SUM-GAN$_{dpp}$ on video # 31 of WCE-2019-Video dataset (F-score=43.3%)



(b) Cycle-SUM on video # 31 of WCE-2019-Video dataset (F-score=46.1%)

(c) SUM-GAN-AAE on video # 31 of WCE-2019-Video dataset (F-score=46.5%)



(d) dppLSTM on video # 31 of WCE-2019-Video dataset (F-score=40.2%)

(e) AALVS on video # 31 of WCE-2019-Video dataset (F-score=47.2%)

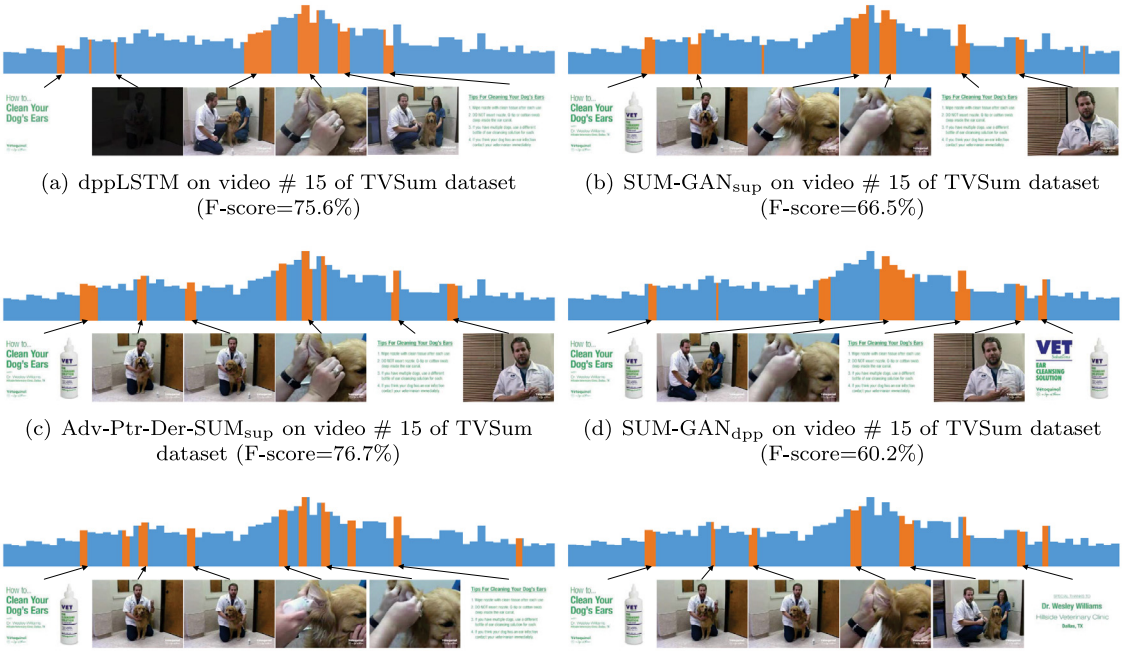
Fig. 7. Exemplar video summaries by other methods. Example summaries from a sample video 31 in WCE-2019-Video. The blue bars show the annotation importance scores. The colored segments are the selected subsets using the specified methods.

are publicly accessible[3,4]. The comparisons in Table 4, between dppLSTM, AALVS, and Adv-Ptr-Der-SUM$_{sup}$, show that supervised variant of our approach significantly outperforms dppLSTM by 9.3%. We believe that outstanding performance is mainly because the pointer network and de-redundancy mechanism can help to produce a good summary, which provide more similar content representation as of the original frame sequences. Furthermore, this variant beats AALVS by 1.2%, which indicates the model may benefit from de-redundancy mechanism and pointer network by different ways.

Comparison with unsupervised approaches on public benchmarks. For fair comparison with other methods and showing the generality of our models, we evaluate the Adv-Ptr-Der-SUM model on two public benchmark datasets: SumMe [56]

[3] https://github.com/kezhang-cs/Video-Summarization-with-LSTM, (last accessed on Apr. 10, 2018).

[4] https://github.com/tsujuifu/pytorch_vsum-ptr-gan, (last accessed on Mar. 21, 2019).



(a) dppLSTM on video # 15 of TVSum dataset (F-score=75.6%)

(b) SUM-GAN$_{sup}$ on video # 15 of TVSum dataset (F-score=66.5%)

(c) Adv-Ptr-Der-SUM$_{sup}$ on video # 15 of TVSum dataset (F-score=76.7%)

(d) SUM-GAN$_{dpp}$ on video # 15 of TVSum dataset (F-score=60.2%)

(e) Cycle-SUM on video # 15 of TVSum dataset (F-score=70.8%)

(f) Adv-Ptr-Der-SUM on video # 15 of TVSum dataset (F-score=71.7%)

Fig. 8. Comparisons of exemplar video summaries by Adv-Ptr-Der-SUM and other methods in the supervised and unsupervised manners. The exemplar video is from TVSum [57]. The blue bars show the ground-truth importance scores. The orange bars are selected subsets of all frames.

and TVSum [57], to verify its effectiveness on user multimedia video summarization. We tentatively empirically conduct a set of experiments and compare our method with other unsupervised approaches. The results of comparison are presented in Table 5.

Table 5 shows the experimental results of Adv-Ptr-Der-SUM against other unsupervised approaches on SumMe and TVSum. As one can see, Adv-Ptr-Der-SUM outperforms most of the existing unsupervised approaches except for CSNet [50] and SUM-GAN-AAE [67] on both datasets by large margins. The performance improve by 1.7% and 0.7% over the third best results, i.e., Cycle-SUM [51] on SumMe and TVSum, respectively. This clearly shows the effectiveness of our proposed approach and well proves that it can be applied to the user multimedia video summarization.

The experimental results show that (1) compared to most of the existing unsupervised approaches except CSNet and SUM-GAN-AAE, our method performs the best on both two datasets. We believe that this is because GAN can be used to unsupervised learning and help to improve the performance; (2) de-redundancy mechanism can reduce the redundancy and highlight diversity information, which may be highly related to the key shots or frames to be selected; (3) in contrast to other techniques, pointer network may more effectively address the issues of variable-range structural dependencies.

Additionally, it can be seen from Tables 4 and 5 that the performance of both SUM-GAN-AAE and Adv-Ptr-Der-SUM on WCE-2019-Video dataset may fall out of step with of both them on two public benchmarks. We believe because it may be that there are different characteristics between both datasets, resulting in different feature representations. Furthermore, the de-redundancy mechanism of Adv-Ptr-Der-SUM is originally intended to eliminate the redundancy in WCE videos, which may be more suitable for WCE, rather than user multimedia videos.

Comparison with supervised approaches on public benchmarks. Similarly, for fairness and showing generality, we evaluate the supervised model on two public benchmarks: SumMe [56] and TVSum [57]. The results are presented in Table 6. As shown in Table 6, it can be seen that the performance of Adv-Ptr-Der-SUM$_{sup}$ is better than other existing supervised approaches, and is even superior than the supervised approach in [66] on TVSum (64.5 vs. 63.6). Also, Adv-Ptr-Der-SUM$_{sup}$ outperforms the mentioned supervised approach except for CSNet [50] on SumMe, which performs slightly better than our approach.

To the best of our knowledge, all existing methods excluding CSNet show that almost all supervised techniques can better improve performance than unsupervised approaches. This experimental result can be consistent with the prior point of view.

4.8. Qualitative results

In this subsection, we first provide qualitative results to better illustrate how well the variations of Adv-Ptr-Der-SUM has selected WCE key frames. Second, we compare our approach with other recent unsupervised methods: SUM-GAN$_{dpp}$ [52], Cycle-SUM [51], and SUM-GAN-AAE [67], as well as supervised dppLSTM [53] and AALVS [66]. The two qualitative results are analyzed on our WCE-2019-Video dataset. Finally, we give an example of comparison results between our approach and other methods on public user video dataset: TVSum.

Fig. 6 shows the selected frames by different variations of our approach on an example video in WCE-2019-Video. The blue background represents the ground-truth importance scores, while the colored regions are the selected subsets by different methods. The illustrations of different variants support the results presented in Table 3.

Fig. 7 demonstrates summarization examples from a sample video #31 in WCE-2019-Video, which generated by the methods including supervised and unsupervised: SUM-GAN$_{dpp}$, Cycle-SUM, and SUM-GAN-AAE, as well as dppLSTM and AALVS. As shown in Figs. 6 and 7, all unsupervised and supervised versions of Adv-Ptr-Der-SUM selects shorter but more key frames with non- or less-redundancy than the other five corresponding models, which

mean that they have higher F-scores than the above-mentioned methods.

In order to verify the effectiveness of our approach on user multimedia video summarization, we conduct several qualitative experiments on video #15 of TVSum [57] dataset by using our approach and others. The comparisons of the results are shown in Fig. 8. It can be seen from Fig. 8 that our approaches including supervised and unsupervised have a relatively high scores on video #15 of TVSum.

5. Conclusion

In this paper, we propose an unsupervised WCE video summarization method termed Adv-Ptr-Der-SUM for frame-level WCE video summarization by integrating long short-term memory (LSTM), variational autoencoder (VAE), generative adversarial network (GAN), pointer network (Ptr-Net), and de-redundancy mechanism (DM) etc. techniques. Extensive experiments on our WCE-2019-Video dataset and two public multimedia benchmarks: SumMe and TVSum, show that our approach can achieve a competitive result on both WCE and user video summarizations in the unsupervised and supervised settings, which well verifies the effectiveness of Adv-Ptr-Der-SUM model. Concretely, using the Adv-Ptr-Der-SUM model, we achieve F-scores of 44.6%, 43.6%, and 58.3% on WCE-2019-Video, two benchmarks: SumMe and TVSum, respectively. Also, using the supervised variant of Adv-Ptr-Der-SUM model, we can achieve F-scores of 45.5%, 47.3%, and 64.5% on the above-mentioned three datasets, respectively.

CRediT authorship contribution statement

Libin Lan: Conceptualization, Methodology, Software, Data curation, Writing - original draft. **Chunxiao Ye:** Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (Grant No. 2017YFB0802400). In addition, we thank Juan Zhou and her colleagues from the Second Affiliated Hospital, Third Military Medical University, for the helpful discussions and suggestions on annotating frame-level importance scores on WCE-2019-Video dataset. We also thank the Chongqing Jinshan Science & Technology (Group) Co., Ltd., for providing vital support with raw WCE videos. We would also like to thank all the authors for their prior excellent work on video summarization, including medical and user multimedia videos. We would also like to thank the anonymous reviewers for their helpful comments which have led to many improvements in this paper.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, Nature 405 (6785) (2000) 417, http://dx.doi.org/10.1038/35013140.

[2] N.M. Lee, G.M. Eisen, 10 years of capsule endoscopy: an update, Expert Rev. Gastroenterol. Hepatol. 4 (4) (2010) 503—512, http://dx.doi.org/10.1586/egh.10.44.

[3] G. Gay, M. Delvaux, J. Rey, The role of video capsule endoscopy in the diagnosis of digestive diseases: a review of current possibilities, Endoscopy 36 (10) (2004) 913—920, http://dx.doi.org/10.1055/s-2004-825868.

[4] D. Hartmann, H. Schmidt, G. Bolz, D. Schilling, F. Kinzel, A. Eickhoff, W. Huschner, K. Möller, R. Jakobs, P. Reitzig, U. Weickert, K. Gellert, H. Schultz, K. Guenther, H. Hollerbuhl, K. Schoenleben, H.-J. Schulz, J.F. Riemann, A prospective two-center study comparing wireless capsule endoscopy with intraoperative enteroscopy in patients with obscure GI bleeding, Gastrointest. Endosc. 61 (7) (2005) 826–832, http://dx.doi.org/10.1016/S0016-5107(05)00372-X.

[5] D.G. Adler, M.A. Knipschield, C.J. Gostout, A prospective comparison of capsule endoscopy and push enteroscopy in patients with GI bleeding of obscure origin, Gastrointest. Endosc. 59 (4) (2004) 492–498, http://dx.doi.org/10.1016/S0016-5107(03)02862-1.

[6] Y. Yuan, B. Li, Q.H. Meng, WCE abnormality detection based on saliency and adaptive locality-constrained linear coding, IEEE Trans. Autom. Sci. Eng. 14 (1) (2017) 149–159, http://dx.doi.org/10.1109/TASE.2016.2610579.

[7] Y. Yuan, B. Li, M.Q. Meng, Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images, IEEE Trans. Autom. Sci. Eng. 13 (2) (2016) 529–535, http://dx.doi.org/10.1109/TASE.2015.2395429.

[8] Y. Yuan, B. Li, M.Q. Meng, Bleeding frame and region detection in the wireless capsule endoscopy video, IEEE J. Biomed. Health Inf. 20 (2) (2016) 624–630, http://dx.doi.org/10.1109/JBHI.2015.2399502.

[9] A. Mata, J. Llach, J. Bordas, Wireless capsule endoscopy, World J. Gastroenterol. 14 (13) (2008) 1969–1971, http://dx.doi.org/10.3748/wjg.14.1969.

[10] M. Pennazio, Capsule endoscopy: Where are we after 6 years of clinical use? Dig. Liver Dis. 38 (12) (2006) 867–878, http://dx.doi.org/10.1016/j.dld.2006.09.007.

[11] A. Koulaouzidis, E. Rondonotti, A. Karargyris, Small-bowel capsule endoscopy: A ten-point contemporary review, World J. Gastroenterol. 19 (24) (2013) 3726–3746, http://dx.doi.org/10.3748/wjg.v19.i24.3726.

[12] A. Koulaouzidis, D.K. Iakovidis, A. Karargyris, J.N. Plevris, Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions, Expert Rev. Gastroenterol. Hepatol. 9 (2) (2015) 217—235, http://dx.doi.org/10.1586/17474124.2014.952281.

[13] D.K. Iakovidis, A. Koulaouzidis, Software for enhanced video capsule endoscopy: challenges for essential progress, Nat. Rev. Gastroenterol. Hepatol. 12 (2015) 172–186, http://dx.doi.org/10.1038/nrgastro.2015.13.

[14] M. Alagappan, J.R.G. Brown, Y. Mori, T.M. Berzin, Artificial intelligence in gastrointestinal endoscopy: The future is almost here, World J. Gastrointest. Endosc. 10 (10) (2018) 239–249, http://dx.doi.org/10.4253/wjge.v10.i10.239.

[15] X. Dray, D. Iakovidis, C. Houdeville, R. Jover, D. Diamantis, A. Histace, A. Koulaouzidis, Artificial intelligence in small bowel capsule endoscopy - current status, challenges and future promise, J. Gastroenterol. Hepatol. 36 (1) (2021) 12–19, http://dx.doi.org/10.1111/jgh.15341.

[16] B. Li, M.Q. Meng, Tumor recognition in wireless capsule endoscopy images using textural features and SVM-based feature selection, IEEE Trans. Inf. Technol. Biomed. 16 (3) (2012) 323–329, http://dx.doi.org/10.1109/TITB.2012.2185807.

[17] Y. Yuan, J. Wang, B. Li, M.Q. Meng, Saliency based ulcer detection for wireless capsule endoscopy diagnosis, IEEE Trans. Med. Imaging 34 (10) (2015) 2046–2057, http://dx.doi.org/10.1109/TMI.2015.2418534.

[18] Y. Yuan, X. Yao, J. Han, L. Guo, M.Q. Meng, Discriminative joint-feature topic model with dual constraints for WCE classification, IEEE Trans. Cybern. 48 (7) (2018) 2074–2085, http://dx.doi.org/10.1109/TCYB.2017.2726818.

[19] A. Karargyris, N. Bourbakis, Detection of small bowel polyps and ulcers in wireless capsule endoscopy videos, IEEE Trans. Biomed. Eng. 58 (10) (2011) 2777–2786, http://dx.doi.org/10.1109/TBME.2011.2155064.

[20] L. Lan, C. Ye, C. Wang, S. Zhou, Deep convolutional neural networks for WCE abnormality detection: Cnn architecture, region proposal and transfer learning, IEEE Access 7 (2019) 30017–30032, http://dx.doi.org/10.1109/ACCESS.2019.2901568.

[21] X. Jia, X. Xing, Y. Yuan, L. Xing, M.Q. Meng, Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition, Proc. IEEE 108 (1) (2020) 178–197, http://dx.doi.org/10.1109/JPROC.2019.2950506.

[22] K. Muhammad, S. Khan, N. Kumar, J. Del Ser, S. Mirjalili, Vision-based personalized wireless capsule endoscopy for smart healthcare: Taxonomy, literature review, opportunities and challenges, Future Gener. Comput. Syst. 113 (2020) 266–280, http://dx.doi.org/10.1016/j.future.2020.06.048.

[23] D.K. Iakovidis, S. Tsevas, A. Polydorou, Reduction of capsule endoscopy reading times by unsupervised image mining, Comput. Med. Imaging Graph. 34 (6) (2010) 471–478, http://dx.doi.org/10.1016/j.compmedimag.2009.11.005.

[24] D.K. Iakovidis, S. Tsevas, D. Maroulis, A. Polydorou, Unsupervised summarisation of capsule endoscopy video, in: 2008 4th International IEEE Conference Intelligent Systems, Vol. 1, 2008, pp. 3–15–3–20, http://dx.doi.org/10.1109/IS.2008.4670414.

[25] S. Tsevas, D. Iakovidis, D. Maroulis, E. Pavlakis, A. Polydorou, Non-negative matrix factorization for endoscopic video summarization, in: Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications, in: SETN '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 425–430, http://dx.doi.org/10.1007/978-3-540-87881-0_44.

[26] A. Biniaz, R.A. Zoroofi, M.R. Sohrabi, Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis, Biomed. Signal Process. Control 59 (2020) 101897, http://dx.doi.org/10.1016/j.bspc.2020.101897.

[27] A. Shiotani, R. Nishi, K. Honda, M. Kawakami, H. Imamura, H. Matsumoto, K. ichi Tarumi, T. Kamada, J. Hata, K. Haruma, W1192 evaluation of quick view of rapid 5 access for examination of video capsule endoscopies, Gastroenterology 138 (5, Supplement 1) (2010) S–671, http://dx.doi.org/10.1016/S0016-5085(10)63086-7, 2010 DDW Abstract Supplement.

[28] U. Günther, S. Daum, M. Zeitz, C. Bojarski, Capsule endoscopy: comparison of two different reading modes, Int. J. Colorectal. Dis. 27 (2012) 521–525, http://dx.doi.org/10.1007/s00384-011-1347-9.

[29] H. Vu, T. Echigo, R. Sagawa, K. Yagi, M. Shiba, K. Higuchi, T. Arakawa, Y. Yagi, Controlling the display of capsule endoscopy video for diagnostic assistance, IEICE Trans. Inf. Syst. E92-D (3) (2009) 512–528, http://dx.doi.org/10.1587/transinf.E92.D.512.

[30] X. Chu, C.K. Poh, L. Li, K.L. Chan, S. Yan, W. Shen, T.M. Htwe, J. Liu, J.H. Lim, E.H. Ong, et al., Epitomized summarization of wireless capsule endoscopic videos for efficient visualization, in: Proceedings of the 13th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part II, in: MICCAI'10, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 522–529, http://dx.doi.org/10.1007/978-3-642-15745-5_64.

[31] E. Spyrou, D. Diamantis, D.K. Iakovidis, Panoramic visual summaries for efficient reading of capsule endoscopy videos, in: 2013 8th International Workshop on Semantic and Social Media Adaptation and Personalization, 2013, pp. 41–46, http://dx.doi.org/10.1109/SMAP.2013.21.

[32] H. Chen, X. Wu, G. Tao, Q. Peng, Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos, Neurocomputing 229 (2017) 77–87, http://dx.doi.org/10.1016/j.neucom.2016.06.077, Advances in computing techniques for big medical image data.

[33] Q. Zhao, G.E. Mullin, M.Q.-H. Meng, T. Dassopoulos, R. Kumar, A general framework for wireless capsule endoscopy study synopsis, Comput. Med. Imaging Graph. 41 (2015) 108–116, http://dx.doi.org/10.1016/j.compmedimag.2014.05.011, Machine Learning in Medical Imaging.

[34] M. Drozdzal, L. Igual, J. Vitrià, C. Malagelada, F. Azpiroz, P. Radeva, Aligning endoluminal scene sequences in wireless capsule endoscopy, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 117–124, http://dx.doi.org/10.1109/CVPRW.2010.5543456.

[35] H. Liu, N. Pan, H. Lu, E. Song, Q. Wang, C.C. Hung, Wireless capsule endoscopy video reduction based on camera motion estimation, J. Digit. Imaging (2013) http://dx.doi.org/10.1007/s10278-012-9519-x.

[36] B. Li, M.Q. Meng, C. Hu, Motion analysis for capsule endoscopy video segmentation, in: 2011 IEEE International Conference on Automation and Logistics (ICAL), 2011, pp. 46–51, http://dx.doi.org/10.1109/ICAL.2011.6024682.

[37] H.-G. Lee, M.-K. Choi, B.-S. Shin, S.-C. Lee, Reducing redundancy in wireless capsule endoscopy videos, Comput. Biol. Med. 43 (6) (2013) 670–682, http://dx.doi.org/10.1016/j.compbiomed.2013.02.009.

[38] B. Sushma, P. Aparna, Summarization of wireless capsule endoscopy video using deep feature matching and motion analysis, IEEE Access 9 (2021) 13691–13703, http://dx.doi.org/10.1109/ACCESS.2020.3044759.

[39] B. Li, M.Q. Meng, Capsule endoscopy video boundary detection, in: 2011 IEEE International Conference on Information and Automation, 2011, pp. 373–378, http://dx.doi.org/10.1109/ICINFA.2011.5949020.

[40] R. Nie, H. Yang, H. Peng, W. Luo, W. Fan, J. Zhang, J. Liao, F. Huang, Y. Xiao, Application of structural similarity analysis of visually salient areas and hierarchical clustering in the screening of similar wireless capsule endoscopic images, 2020, arXiv e-prints http://arxiv.org/abs/2004.02805.

[41] R. Sharma, R. Bhadu, S.K. Soni, N. Varma, Reduction of redundant frames in active wireless capsule endoscopy, in: V. Nath, J.K. Mandal (Eds.), Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017), Springer Singapore, Singapore, 2019, pp. 1–7, http://dx.doi.org/10.1007/978-981-10-8234-4_1.

[42] Q. Al-shebani, P. Premaratne, D.J. McAndrew, P.J. Vial, S. Abey, A frame reduction system based on a color structural similarity (CSS) method and bayer images analysis for capsule endoscopy, Artif. Intell. Med. 94 (2019) 18–27, http://dx.doi.org/10.1016/j.artmed.2018.12.008.

[43] Q. Wang, N. Pan, W. Xiong, H. Lu, N. Li, X. Zou, Reduction of bubble-like frames using a RSS filter in wireless capsule endoscopy video, Opt. Laser Technol. 110 (2019) 152–157, http://dx.doi.org/10.1016/j.optlastec.2018.08.051, Special Issue: Optical Imaging for Extreme Environment.

[44] Y. Yuan, M.Q. Meng, Hierarchical key frames extraction for WCE video, in: 2013 IEEE International Conference on Mechatronics and Automation, 2013, pp. 225–229, http://dx.doi.org/10.1109/ICMA.2013.6617922.

[45] J. Chen, Y. Zou, Y. Wang, Wireless capsule endoscopy video summarization: A learning approach based on siamese neural network and support vector machine, in: 2016 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 1303–1308, http://dx.doi.org/10.1109/ICPR.2016.7899817.

[46] J. Chen, Y. Wang, Y.X. Zou, An adaptive redundant image elimination for wireless capsule endoscopy review based on temporal correlation and color-texture feature similarity, in: 2015 IEEE International Conference on Digital Signal Processing (DSP), 2015, pp. 735–739, http://dx.doi.org/10.1109/ICDSP.2015.7251973.

[47] K. Schoeffmann, M.D. Fabro, T. Szkaliczki, A. Böszörmenyi, J. Keckstein, Keyframe extraction in endoscopic video, Multimedia Tools Appl. 74 (2015) 11187–11206, http://dx.doi.org/10.1007/s11042-014-2224-7.

[48] A.F. Smeaton, P. Over, A.R. Doherty, Video shot boundary detection: Seven years of trecvid activity, Comput. Vis. Image Underst. 114 (4) (2010) 411–418, http://dx.doi.org/10.1016/j.cviu.2009.03.011.

[49] Z. Ji, K. Xiong, Y. Pang, X. Li, Video summarization with attention-based encoder-decoder networks, IEEE Trans. Circuits Syst. Video Technol. (2019) 1, http://dx.doi.org/10.1109/TCSVT.2019.2904996.

[50] Y. Jung, D. Cho, D. Kim, S. Woo, I.S. Kweon, Discriminative feature learning for unsupervised video summarization, 2018, CoRR abs/1811.09791 http://arxiv.org/abs/1811.09791.

[51] L. Yuan, F.E.H. Tay, P. Li, L. Zhou, J. Feng, Cycle-SUM: Cycle-consistent adversarial LSTM networks for unsupervised video summarization, 2019, CoRR abs/1904.08265 http://arxiv.org/abs/1904.08265.

[52] B. Mahasseni, M. Lam, S. Todorovic, Unsupervised video summarization with adversarial LSTM networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2982–2991, http://dx.doi.org/10.1109/CVPR.2017.318.

[53] K. Zhang, W. Chao, F. Sha, K. Grauman, Video summarization with long short-term memory, 2016, CoRR abs/1605.08110 http://arxiv.org/abs/1605.08110.

[54] S. hua Zhong, J. Wu, J. Jiang, Video summarization via spatio-temporal deep architecture, Neurocomputing 332 (2019) 224–235, http://dx.doi.org/10.1016/j.neucom.2018.12.040.

[55] Z. Ji, Y. Zhang, Y. Pang, X. Li, J. Pan, Multi-video summarization with query-dependent weighted archetypal analysis, Neurocomputing 332 (2019) 406–416, http://dx.doi.org/10.1016/j.neucom.2018.12.038.

[56] M. Gygli, H. Grabner, H. Riemenschneider, L. Van Gool, Creating summaries from user videos, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 505–520, http://dx.doi.org/10.1007/978-3-319-10584-0_33.

[57] Y. Song, J. Vallmitjana, A. Stent, A. Jaimes, Tvsum: Summarizing web videos using titles, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 5179–5187, http://dx.doi.org/10.1109/CVPR.2015.7299154.

[58] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[59] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, 2013, http://arxiv.org/abs/1312.6114.

[60] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, http://arxiv.org/abs/1409.0473.

[61] M. Luong, H. Pham, C.D. Manning, Effective approaches to attention-based neural machine translation, 2015, CoRR abs/1508.04025 http://arxiv.org/abs/1508.04025.

[62] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014, http://arxiv.org/abs/1406.2661.

[63] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using LSTMs, 2015, http://arxiv.org/abs/1502.04681.

[64] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, B. Guo, Unsupervised extraction of video highlights via robust recurrent auto-encoders, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, 2015, http://dx.doi.org/10.1109/iccv.2015.526.

[65] L. Lebron Casas, E. Koblents, Video summarization with LSTM and deep attention models, in: I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, S. Vrochidis (Eds.), MultiMedia Modeling, Springer International Publishing, Cham, 2019, pp. 67–79, http://dx.doi.org/10.1007/978-3-030-05716-9_6.

[66] T. Fu, S. Tai, H. Chen, Attentive and adversarial learning for video summarization, in: IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019, 2019, pp. 1579–1587, http://dx.doi.org/10.1109/WACV.2019.00173.

[67] E.E. Apostolidis, E. Adamantidou, A.I. Metsai, V. Mezaris, I. Patras, Unsupervised video summarization via attention-driven adversarial learning, in: MultiMedia Modeling - 26th International Conference, MMM 2020, Daejeon, South Korea, January 5-8, 2020, Proceedings, Part I, in: Lecture Notes in Computer Science, Vol. 11961, Springer, 2020, pp. 492–504, http://dx.doi.org/10.1007/978-3-030-37731-1_40.

[68] B. Zhao, X. Li, X. Lu, TTH-RNN: Tensor-train hierarchical recurrent neural network for video summarization, IEEE Trans. Ind. Electron. (2020) 1, http://dx.doi.org/10.1109/TIE.2020.2979573.

[69] B. Zhao, X. Li, X. Lu, Property-constrained dual learning for video summarization, IEEE Trans. Neural Netw. Learn. Syst. (2019) 1–12, http://dx.doi.org/10.1109/TNNLS.2019.2951680.

[70] Z. Ji, Y. Zhao, Y. Pang, X. Li, J. Han, Deep attentive video summarization with distribution consistency learning, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–11, http://dx.doi.org/10.1109/TNNLS.2020.2991083.

[71] Z. Ji, F. Jiao, Y. Pang, L. Shao, Deep attentive and semantic preserving video summarization, Neurocomputing 405 (2020) 200–207, http://dx.doi.org/10.1016/j.neucom.2020.04.132.

[72] Z. Tu, Z. Lu, Y. Liu, X. Liu, H. Li, Modeling coverage for neural machine translation, 2016, http://arxiv.org/abs/1601.04811.

[73] O. Vinyals, M. Fortunato, N. Jaitly, Pointer networks, 2015, http://arxiv.org/abs/1506.03134.

[74] J. Gu, Z. Lu, H. Li, V.O.K. Li, Incorporating copying mechanism in sequence-to-sequence learning, 2016, http://arxiv.org/abs/1603.06393.

[75] A. See, P.J. Liu, C.D. Manning, Get to the point: Summarization with pointer-generator networks, 2017, http://arxiv.org/abs/1704.04368.

[76] F. Sun, P. Jiang, H. Sun, C. Pei, W. Ou, X. Wang, Multi-source pointer network for product title summarization, 2018, http://arxiv.org/abs/1808.06885.

[77] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, Y. Bengio, Pointing the unknown words, 2016, http://arxiv.org/abs/1603.08148.

[78] L. Zhu, Z. Xu, Y. Yang, Bidirectional multirate reconstruction for temporal modeling in videos, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1339–1348, http://dx.doi.org/10.1109/CVPR.2017.147.

[79] L. Zhu, H. Fan, Y. Luo, M. Xu, Y. Yang, Temporal cross-layer correlation mining for action recognition, IEEE Trans. Multimed. (2021) 1, http://dx.doi.org/10.1109/TMM.2021.3057503.

[80] B. Gong, W.-L. Chao, K. Grauman, F. Sha, Diverse sequential subset selection for supervised video summarization, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, in: NIPS'14, MIT Press, Cambridge, MA, USA, 2014, pp. 2069–2077.

[81] K. Zhang, K. Grauman, F. Sha, Retrospective encoders for video summarization, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 391–408, http://dx.doi.org/10.1007/978-3-030-01237-3_24.

[82] X. Zhu, J. Fan, A.K. Elmagarmid, W.G. Aref, Hierarchical video summarization for medical data, in: Storage and Retrieval for Media Databases 2002, Vol. 4676, SPIE, 2001, pp. 395–406, http://dx.doi.org/10.1117/12.451110.

[83] P.D. Byrnes, W.E. Higgins, Efficient bronchoscopic video summarization, IEEE Trans. Biomed. Eng. 66 (3) (2019) 848–863, http://dx.doi.org/10.1109/TBME.2018.2859322.

[84] Y. Shen, P. Guturu, B.P. Buckles, Wireless capsule endoscopy video segmentation using an unsupervised learning approach based on probabilistic latent semantic analysis with scale invariant features, IEEE Trans. Inf. Technol. Biomed. 16 (1) (2012) 98–105, http://dx.doi.org/10.1109/TITB.2011.2171977.

[85] Y. Fu, H. Liu, Y. Cheng, T. Yan, T. Li, M.Q.. Meng, Key-frame selection in WCE video based on shot detection, in: Proceedings of the 10th World Congress on Intelligent Control and Automation, 2012, pp. 5030–5034, http://dx.doi.org/10.1109/WCICA.2012.6359431.

[86] J.S. Huo, Y.X. Zou, L. Li, An advanced WCE video summary using relation matrix rank, in: Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics, 2012, pp. 675–678, http://dx.doi.org/10.1109/BHI.2012.6211673.

[87] C. Loukas, C. Varytimidis, K. Rapantzikos, M.A. Kanakis, Keyframe extraction from laparoscopic videos based on visual saliency detection, Comput. Methods Programs Biomed. 165 (2018) 13–23, http://dx.doi.org/10.1016/j.cmpb.2018.07.004.

[88] R. Hamza, K. Muhammad, Z. Lv, F. Titouna, Secure video summarization framework for personalized wireless capsule endoscopy, Pervasive Mob. Comput. 41 (2017) 436–450, http://dx.doi.org/10.1016/j.pmcj.2017.03.011.

[89] B. Li, M.Q. Meng, Q. Zhao, Wireless capsule endoscopy video summary, in: 2010 IEEE International Conference on Robotics and Biomimetics, 2010, pp. 454–459, http://dx.doi.org/10.1109/ROBIO.2010.5723369.

[90] G. Gallo, E. Granata, A. Torrisi, Information theory based WCE video summarization, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 4198–4201, http://dx.doi.org/10.1109/ICPR.2010.1020.

[91] Y.-F. Ma, L. Lu, H.-J. Zhang, M. Li, A user attention model for video summarization, in: Proceedings of the Tenth ACM International Conference on Multimedia, in: MULTIMEDIA '02, Association for Computing Machinery, New York, NY, USA, 2002, pp. 533–542, http://dx.doi.org/10.1145/641007.641116.

[92] F. Altaf, S.M.S. Islam, N. Akhtar, N.K. Janjua, Going deep in medical image analysis: Concepts, methods, challenges, and future directions, IEEE Access 7 (2019) 99540–99572, http://dx.doi.org/10.1109/access.2019.2929365.

[93] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. van der Laak, B. van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, http://dx.doi.org/10.1016/j.media.2017.07.005.

[94] G.H.-J. Kwak, P. Hui, Deephealth: Deep learning for health informatics, 2019, http://arxiv.org/abs/1909.00384.

[95] X. Yi, E. Walia, P. Babyn, Generative adversarial network in medical imaging: A review, Med. Image Anal. 58 (2019) 101552, http://dx.doi.org/10.1016/j.media.2019.101552.

[96] J.M. Wolterink, K. Kamnitsas, C. Ledig, I. Išgum, Generative adversarial networks and adversarial methods in biomedical image analysis, 2018, http://arxiv.org/abs/1810.10352.

[97] Y. Tian, S. Fu, A descriptive framework for the field of deep learning applications in medical images, Knowl.-Based Syst. 210 (2020) 106445, http://dx.doi.org/10.1016/j.knosys.2020.106445.

[98] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, Y. Barash, Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis, Gastrointest. Endosc. 92 (4) (2020) 831 – 839.e8, http://dx.doi.org/10.1016/j.gie.2020.04.039.

[99] S. Tsevas, D.K. Iakovidis, D. Maroulis, E. Pavlakis, Automatic frame reduction of wireless capsule endoscopy video, in: 2008 8th IEEE International Conference on BioInformatics and BioEngineering, 2008, pp. 1–6, http://dx.doi.org/10.1109/BIBE.2008.4696805.

[100] Z. Sun, B. Li, R. Zhou, H. Zheng, M.Q. Meng, Removal of non-informative frames for wireless capsule endoscopy video segmentation, in: 2012 IEEE International Conference on Automation and Logistics, 2012, pp. 294–299, http://dx.doi.org/10.1109/ICAL.2012.6308214.

[101] M.M.B. Ismail, O. Bchir, A.Z. Emam, Endoscopy video summarization based on unsupervised learning and feature discrimination, in: 2013 Visual Communications and Image Processing (VCIP), 2013, pp. 1–6, http://dx.doi.org/10.1109/VCIP.2013.6706410.

[102] B. Münzer, K. Schoeffmann, L. Böszörmenyi, Domain-specific video compression for long-term archiving of endoscopic surgery videos, in: 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS), 2016, pp. 312–317, http://dx.doi.org/10.1109/CBMS.2016.28.

[103] U. von Öhsen, J.M. Marcinczak, A.F.M. Vélez, R.-R. Grigat, Keyframe selection for robust pose estimation in laparoscopic videos, in: D.R.H. I.I.I., K.H. Wong (Eds.), Medical Imaging 2012: Image-Guided Procedures, Robotic Interventions, and Modeling, Vol. 8316, International Society for Optics and Photonics, SPIE, 2012, pp. 306–313, http://dx.doi.org/10.1117/12.911381.

[104] S. Wang, Y. Cong, J. Cao, Y. Yang, Y. Tang, H. Zhao, H. Yu, Scalable gastroscopic video summarization via similar-inhibition dictionary selection, Artif. Intell. Med. 66 (2016) 1–13, http://dx.doi.org/10.1016/j.artmed.2015.08.006.

[105] C. Li, A.B. Hamza, N. Bouguila, X. Wang, F. Ming, G. Xiao, Online redundant image elimination and its application to wireless capsule endoscopy, Signal Imag. Video Process. 8 (8) (2014) 1497–1506, http://dx.doi.org/10.1007/s11760-012-0384-3.

[106] Y. Miao, P. Blunsom, Language as a latent variable: Discrete generative models for sentence compression, 2016, http://arxiv.org/abs/1609.07317.

[107] A.B.L. Larsen, S.K. Sønderby, O. Winther, Autoencoding beyond pixels using a learned similarity metric, 2015, CoRR abs/1512.09300 http://arxiv.org/abs/1512.09300.

[108] J.J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, 2016, CoRR abs/1609.03126 arXiv:1609.03126.

[109] M. Keuchel, F. Hagenmüller, H. Tajiri, Video Capsule Endoscopy: A Reference Guide and Atlas, Springer-Verlag, Berlin, Heidelberg, 2014, http://dx.doi.org/10.1007/978-3-662-44062-9.

[110] X. Li, B. Zhao, X. Lu, A general framework for edited video and raw video summarization, IEEE Trans. Image Process. 26 (8) (2017) 3652–3664, http://dx.doi.org/10.1109/TIP.2017.2695887.

[111] M. Gygli, H. Grabner, L. Van Gool, Video summarization by learning submodular mixtures of objectives, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3090–3098, http://dx.doi.org/10.1109/CVPR.2015.7298928.

[112] K. Zhou, Y. Qiao, T. Xiang, Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward, 2017, http://arxiv.org/abs/1801.00054.

[113] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, http://arxiv.org/abs/1412.6980.

[114] A. Paszke, am Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, pp. 1–4.

[115] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 2014, http://arxiv.org/abs/1409.4842.

[116] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, http://dx.doi.org/10.1007/s11263-015-0816-y.