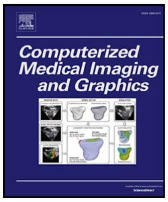
Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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



Anatomical landmarks localization for capsule endoscopy studies

Pablo Laiz [a,*], Jordi Vitrià [a], Pere Gilabert [a], Hagen Wenzek [b], Carolina Malagelada [c], Angus J.M. Watson [d], Santi Seguí [a]

[a] Department of Mathematics and Computer Science, Universitat de Barcelona, Barcelona, Spain
[b] CorporateHealth International ApS, Denmark
[c] Digestive System Research Unit, University Hospital Vall d'Hebron, Spain
[d] Department of Surgery, Raigmore Hospital, Inverness, UK

ARTICLE INFO

Keywords:
Anatomical landmarks
Deep learning
Wireless capsule endoscopy
Organ detection

ABSTRACT

Wireless Capsule Endoscopy is a medical procedure that uses a small, wireless camera to capture images of the inside of the digestive tract. The identification of the entrance and exit of the small bowel and of the large intestine is one of the first tasks that need to be accomplished to read a video. This paper addresses the design of a clinical decision support tool to detect these anatomical landmarks. We have developed a system based on deep learning that combines images, timestamps, and motion data to achieve state-of-the-art results. Our method does not only classify the images as being inside or outside the studied organs, but it is also able to identify the entrance and exit frames. The experiments performed with three different datasets (one public and two private) show that our system is able to approximate the landmarks while achieving high accuracy on the classification problem (inside/outside of the organ). When comparing the entrance and exit of the studied organs, the distance between predicted and real landmarks is reduced from 1.5 to 10 times with respect to previous state-of-the-art methods.

1. Introduction

Wireless Capsule Endoscopy (WCE) (Iddan et al., 2000) is a medical procedure designed to visualize the entire digestive tract through a swallowed vitamin-size capsule, which is propelled by peristalsis via the esophagus, stomach, small intestine, and large intestine (also referred to as colon). WCE offers several benefits to patients, clinicians, and the healthcare system in comparison with traditional endoscopic procedures. It does not require sedation, is less likely to cause discomfort, and presents fewer potential complications. It also minimizes the needed medical resources compared to the standard screening technique (Darrow, 2014).

Currently, in several countries, small bowel WCE is used as the first indication for obscure gastrointestinal (GI) bleeding, Crohn's disease, and to a lesser extent, screening in polyposis syndromes, celiac disease, or other small bowel pathologies (Trasolini and Byrne, 2021). Meanwhile, colon WCE is increasingly recognized as a reliable option for polyp detection, investigation of inflammatory bowel diseases or completion of an incomplete colonoscopy (Yung et al., 2016; Koulaouzidis et al., 2021).

Unfortunately, the adoption of this technique is below the initial expectation, mainly because WCE: (1) does not admit any surgical intervention; (2) does not provide the exact location of the pathology or organs; and (3) generates recordings with thousands of frames that must be reviewed by experts, entailing a complex and time-consuming task. Even an experienced reader may require at least an hour to analyze the data of a single patient (Maieron et al., 2004; Dokoutsidou et al., 2011; Rondonotti et al., 2020).

Artificial Intelligence (AI) methods are being employed in several solutions to overcome WCE limitations and accelerate the reviewing process for readers. While most studies have been centered on detecting images with abnormalities, such as polyps, tumors, bleeding, or ulcers, few of them are focused on localizing the findings or the anatomical landmarks.

In the clinical field, the localization of anatomical landmarks and abnormalities represents a problem of particular interest as it is essential to guide gastroenterologists during the screening and to take clinical decisions (Iakovidis and Koulaouzidis, 2015). Indeed, the localization of these landmarks is one of the first tasks carried out by the readers and is required to perform a complete exploration (Koulaouzidis et al., 2021).

In this paper, we propose a deep learning method for automatically localizing relevant anatomical landmarks to be used in the clinical routine with different capsule endoscopy devices. The aim of this

[*] Corresponding author.
E-mail address: laizpablo@ub.edu (P. Laiz).

https://doi.org/10.1016/j.compmedimag.2023.102243
Received 14 September 2022; Received in revised form 25 February 2023; Accepted 5 May 2023
Available online 18 May 2023

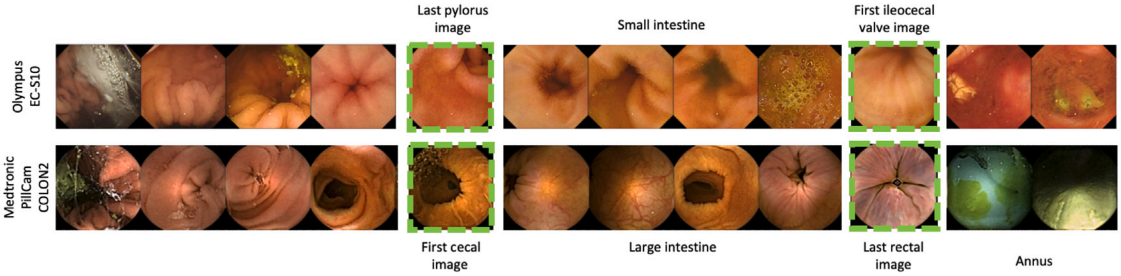


Fig. 1. Illustration of random frames from two GI tracts. The first sample is recorded with the Olympus EC-S10, whereas the second one is obtained with Medtronic PillCam COLON2. The corresponding landmarks of the small bowel (first sample) and the large intestine (second sample) are bordered by a green dashed line.

work is to reduce the average time required to complete the clinical routine which typically takes approximately 25 minutes by a specialist reader (Iakovidis and Koulaouzidis, 2015). To reach this purpose, the method focuses on detecting the end of the pylorus and the ileocecal valve, which delimit the small bowel. For the large intestine, the points of interest are the first cecal and last rectal images. Moreover, the last rectal image ensures the proper identification of the farthest point of capsule progression. These landmarks are illustrated in Fig. 1 bordered by a green dashed line. Each one of them can be visualized in multiple depending on the orientation of the capsule. Though, in some cases, they can be hidden by GI content, which increases the complexity of the task.

First, our system aims to identify all the images between the landmarks using video frames, and additionally, timestamps and motion information that have not been employed in previous studies. Subsequently, the model recognizes the first and last image belonging to the studied organ. The obtained results show that, by providing extra knowledge to the network, the performance of the system increases compared to the state-of-the-art methods and may reduce the average time to complete the clinical routine.

The paper is organized as follows: initially, an overview of the related work in the field is given. Then, our method is presented in detail explaining the key steps, followed by the experimental setup, where the three used databases and metrics are introduced. After that, the results of the experiments are extensively exposed in a quantitative and qualitative manner to prove the performance of the method. Finally, the conclusions and future work are discussed.

2. Related work

The related work can be divided into two main categories: *ad hoc* techniques and deep learning models. First, the traditional statistical methods and machine learning techniques are reviewed. Then, deep learning methods focused on organ classification are summarized.

Berens et al. (2005) were the first to propose a solution for the detection of anatomical landmarks. They employed hue saturation chromaticity histograms to distinguish the stomach, intestine, and colon tissues. Lee et al. (2007) made use of intestinal contractions to locate the boundaries of the organs or unusual events such as intestinal juices, bleeding, and rare capsule movements. Mackiewicz et al. (2008) described the use of color image analysis to discriminate between the esophagus, stomach, small intestine, and colon. Haji-Maghsoudi et al. (2012) proposed an algorithm to classify the same organs using static and non-static features. Li et al. (2015) reported a method that draws a dissimilarity curve implementing the color feature to locate the boundaries between the stomach, small intestine, and large intestine. In these methods, the system performance is assessed as the frame distance (error) between the point in the video where the boundary was manually annotated by a clinician and the one selected by the algorithm.

The latest methods to discriminate between organs are based on deep learning techniques. Zou et al. (2015) proposed a network called DCNN-WCE-CS to classify the digestive organs from WCE images by recognizing high-level semantic features. The network was built with three convolutional layers and a dense layer to classify. Chen et al. (2017) presented two different systems, O-CNN and TO-CNN. The former consisted of a standard Convolutional Neural Network (CNN), "AlexNet", whereas the latter additionally integrated temporal information employing Hidden Markov models. Adewole et al. (2020) compared four state-of-the-art Deep Neural Networks (DNN) to detect the anatomical parts within the GI tract. Zhao et al. (2021) designed a three stages method to detect the boundaries of the small bowel. The method explores long-range temporal dependency with a transformer module, which captures the temporal inter-frame dependencies in short sequences. To locate the starting and ending of the organ, a search algorithm is applied. Finally, Son et al. (2022) proposed a system based on a DNN with temporal filtering (a combination of median and Savitzky–Golay filters) on the predicted probabilities. To detect the boundaries, the method considers the minimum and maximum frame index predicted as small bowel. Although in terms of classification, deep learning methods outperform the obtained results with the extraction of handcrafted features, only Zhao et al. (2021) and Son et al. (2022) apply thresholding techniques to identify the boundaries of the small bowel. To the best of our knowledge, all the studies were performed using private datasets and with only one type of capsule.

3. Method

Our method aims to localize the anatomical landmarks from WCE videos. An overview of the employed strategy is illustrated in Fig. 2. To achieve the primary purpose, the main steps are: (1) Develop a deep learning model to predict the probability of each image to belong to the area of interest, the small bowel or the large intestine; (2) Smooth and mitigate any noisy behavior of the probabilities with extra information (temporal and motion data); (3) Predict the boundaries using a rectangular pulse function by a minimization problem.

3.1. Step 1: Probability prediction

Let $x_i \in X$ be an image, where x_i is the ith-frame of a WCE video X, and $f(\cdot)$, a DNN architecture. The low representation of the image x_i is defined as $x_i' = f(x_i) \in \mathbb{R}^{2048}$. The vector x_i' is extended by adding a new feature containing the temporal information of the frame, $z_i = x_i' \parallel t_i \in \mathbb{R}^{2049}$, $t_i \in [0, 1]$.

The added time-related feature, t_i, is based on the image timestamp and exists for each i. Each image is mapped to a value between zero and one according to:

$$t_i = \frac{\text{timestamp}_i}{\text{video length}} \tag{1}$$

where $timestamp_i$ represents the time (in seconds) of the ith-frame in the video. This equation normalizes all the video lengths and provides the temporal position with respect to the entire video.

The WCE advances through the GI tract recording all organs in a continuous manner. It is worth remarking that although the camera

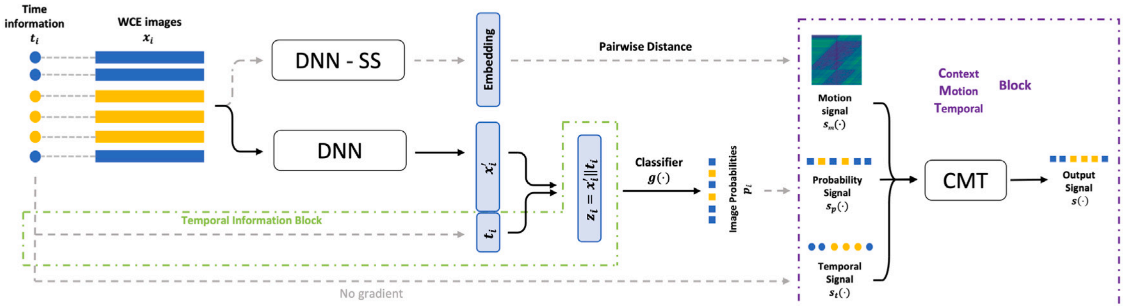


Fig. 2. Overview of the proposed system. The input of the network consists of a sequence of images and their temporal information. The main architecture is a DNN concatenated with the temporal and CMT blocks. The output of the model is a smooth signal with low noise.

might go back and forth, it remains in the same organ. This allows the model to create a relationship between time and organs. The temporal feature added by our system allows the model to discard erroneous predictions in different sections of the video.

Then, using a linear classifier $g(\cdot)$ and the extended vector z_i, the probability p_i of each frame is inferred:

$$p_i = g(z_i) = g(x_i' \parallel t_i) = g(f(x_i) \parallel t_i) \tag{2}$$

3.2. Step 2: Smoothing the probabilities

Fig. 1 contains some examples of frames where the mucosa of the digestive tract is hidden by noisy content (Chen et al., 2017) like bile, bubbles, residues, and liquids. In those frames, the network may yield senseless probabilities. To mitigate this undesirable behavior, it is important not only to properly analyze a still image but the entire sequence. Furthermore, if the context analysis is complemented with the movements of the capsule within the intestine and the temporal information, the developed model can further decrease this erratic behavior.

Given all the frames from a video, the capsule movement signal $s_m(\cdot)$ is obtained by estimating the distance between frames. To calculate them, the time-based self-supervised network, $f_{ss}(\cdot)$, proposed by Pascual et al. (2022) is employed to obtain the embedding for each image x_i, $e_i = f_{ss}(x_i)$. The network generates similar representations for images that are close in time, i.e., consecutive frames from the same sequence have similar embeddings. For images from different sequences, the network yields distant image representations. Then, the Euclidean pairwise distance is computed between the embeddings to obtain the matrix M. The values of M are an approximation of the motion between two frames. Small values of the matrix are caused by small movements of the capsule, while high values mean the opposite.

The visualization of this matrix shows contraction patterns of the GI tract and can suggest where the camera might be located. Because of the length of the video, the complete matrix is difficult to visualize. Therefore, it is simplified as a figure containing the ith-frame, centered in the middle of each row and their 500 nearest temporal neighbors, all of them represented as pixels. The color of each one is the distance between the frame and the central one. The darkest points correspond to small distances, implying that the capsule hardly moves. While, the lighter pixels point out larger distances, which entails a drastic movement of the capsule. Fig. 3 contains three samples of the capsule movement codified as an image. Each one is shown in four parts: the beginning of the video, the first landmark, a random segment of the organ, and the second landmark. The frames containing the landmarks annotated by the experts are represented with black dashed lines.

The sequential analysis is performed in the context-motion-temporal (CMT) block, which smooths the probabilities of those frames with senseless values by combining neighborhood probabilities, motion, and time information. The use of the CMT block is a paradigm shift which works with probabilities and information from the whole video encoded in three signals:

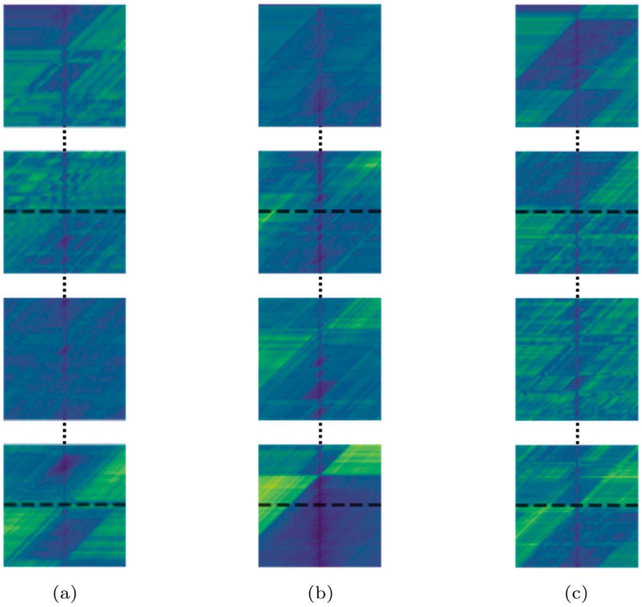


Fig. 3. Movement visualization of the capsule in different dataset videos: (a) *Kvasir-Capsule* dataset, (b) *VH* dataset and, (c) *Capri* dataset. Each row corresponds to the beginning of the video, the first landmark, a random part of the organ, and the second landmark. Inside each patch, the x-axis represents the relationship of the central frame to the other frames, and the y-axis contains the frames in chronological order. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Probability signal, $s_p(\cdot)$, is obtained by concatenating the probabilities inferred in each frame of a video:

$$s_p(i) = g(f(x_i) \parallel t_i), \quad \forall i \mid x_i \in X \tag{3}$$

- Motion signal, $s_m(\cdot)$, is obtained by using the normalized ith-row of the matrix M.
- Temporal signal, $s_t(\cdot)$, is obtained by concatenating the time information of each frame of a video:

$$s_t(i) = t_i, \quad \forall t_i \tag{4}$$

These signals are concatenated vertically to generate a matrix of size $3 \times video\ length$. To calculate the output signal, $s(\cdot)$, a small network called $\mathrm{CMT}_w(\cdot)$ is used. It is composed of two layers of bidirectional LSTM cells and one dense layer over w consecutive frames. This is formalized as:

$$s = \mathrm{CMT}_w(s_p \parallel s_m \parallel s_t) \tag{5}$$

The window size hyper-parameter, w, is a natural odd number that must be determined to achieve optimal results. The overview of this block can be seen in Fig. 4.

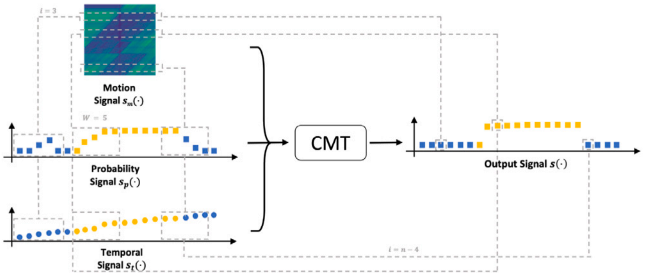


Fig. 4. Overview of the proposed CMT block with $w = 5$. The input of the block is the different signals extracted from processing a WCE video: the probability signal s_p, the motion signal m and the temporal signal s_t. The output signal s is obtained after combining the given information.

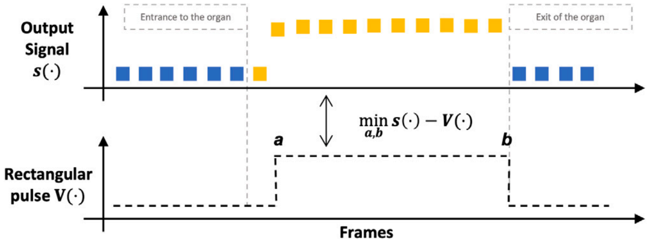


Fig. 5. Overview of the minimization problem, given the output signal of the video and the rectangular pulse function required to solve Eq. (8). Gray lines correspond to the anatomical landmark annotated by the expert.

3.3. Step 3: Boundaries prediction

Finally, a simple but efficient technique is employed to identify the landmarks of the WCE video using the inferred probabilities of each image belonging to an organ. A minimization problem, $\varphi(\cdot)$, is solved over the output signal to identify the boundaries of the organ, as it is shown in Fig. 5. Let $V(t)$ be the rectangular pulse function:

$$V(t) = u(t-a) - u(t-b) \tag{6}$$

where $u(t)$ is the unit step function defined as:

$$u(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases} \tag{7}$$

and a and b are the limits where the function $V(t)$ has value one. To identify the first and last frame of the organ, the distance between the output signal $s(i)$ and the rectangular pulse $V(t)$ is minimized by finding the best a and b values:

$$\begin{aligned} \underset{a,b}{\text{minimize}} \; & \sum_{i=0}^{a-1} s(i) + \sum_{i=a}^{b-1} 1 - s(i) + \sum_{i=b}^{n} s(i) \\ \text{s.t.} \quad & a < b \end{aligned} \tag{8}$$

The optimization of the network weights is carried out using the binary cross-entropy loss to minimize Eqs. (2) and (5). In both cases, the real binary labels inside/outside the organ have been used to compute the cross-entropy during training.

4. Experimental setup

4.1. Datasets

The proposed system is evaluated with one public dataset (*Kvasir-Capsule*) and two private ones (*VH* and *Capri*).

Table 1
Overview of the records in the three datasets used in this paper. The column *#Inside* refers to those frames that are between the landmarks specified in each dataset. Respectively, column *#Outside* refers to the number of frames that do not belong to the area of interest.

Dataset	Partitions	#Patients	#Inside	#Outside	Total
Kvasir-Capsule	Fold 0	12	400K	160K	560K
	Fold 1	12	384K	97K	481K
	Total	**24**	**784K**	**257K**	**1M**
VH	Fold 0	24	602K	246K	848K
	Fold 1	24	592K	249K	841K
	Total	**48**	**1.2M**	**495K**	**1.6M**
Capri	Fold 0	34	347K	148K	495K
	Fold 1	34	393K	97K	490K
	Total	**68**	**740K**	**245K**	**985K**

4.1.1. Kvasir-capsule dataset

This public dataset was collected from 117 examinations at a Norwegian Hospital employing the Olympus Endocapsule 10 System (EC-S10) (Smedsrud et al., 2021). In our case, we only used the set of 24 videos that contains anatomical landmarks of the small bowel, specifically the pylorus and the ileocecal valve. The number of frames per video is $44K$ on average. Small bowel images represent 75.14% of the dataset. However, this dataset does not contain the temporal information of the videos.

4.1.2. VH dataset

The second dataset was obtained from 48 healthy volunteers. Physicians from Vall d'Hebron hospital in Barcelona recorded all the videos using Medtronic PillCam SB3 and labeled the limits of the small bowel. The average number of frames per video is $35K$ with a mean video duration of 04:36:06. The frames between the pylorus and the ileocecal valve represent the 70.68% of this dataset.

4.1.3. Capri dataset

The last used database is composed of 68 colon studies from different patients. All these WCE videos were recorded with Medtronic PillCam COLON2 on behalf of the NHS Highland Raigmore Hospital in Inverness. Images from both cameras, frontal and rear, from the PillCam COLON2 are used in the experiments. The mean duration of the videos is 08:19:51 with an average of $14K$ frames. The colon images represent the 74.63% of the dataset.

4.2. Evaluation criteria

Models are evaluated with a two-fold stratified cross-validation strategy, following the instructions established by Smedsrud et al. (2021). It is worth remarking that the stratified partitions are not based on individual frames but on individual patients. Hence, images from the same patients do not belong to different sets. Table 1 contains the details about each fold for each one of the used datasets.

As in previous (Zou et al., 2015; Chen et al., 2017; Adewole et al., 2020; Zhao et al., 2021; Son et al., 2022), the performance of the method in the classification task is measured with the following metrics: the Area Under the ROC Curve (AUC), Accuracy (ACC), Mean Accuracy (MACC), Specificity (SPEC), and Sensitivity (SENS).

The AUC and MACC are the most appropriate metrics for evaluating the performance of a binary classification model on imbalanced datasets. The AUC measures the model's ability to distinguish between images that belong to the target organ and those that do not, while the MACC and ACC measure the number of images that are correctly predicted. It is important to note that relying solely on SENS and SPEC rates for comparison can be problematic, as these can vary depending on the chosen cut-off thresholds.

Table 2
Window size hyper-parameters tested during training. The metrics used to identify which is the best value are the AUC and the total median error obtained in a two-fold cross-validation.

Window size	Datasets											
	Kvasir-Capsule				VH				Capri			
	AUC	Median error			AUC	Median error			AUC	Median error		
		Entrance	Exit	Total		Entrance	Exit	Total		Entrance	Exit	Total
11	95.66	58.25	982.00	1040.25	98.09	46.50	266.75	313.25	99.61	4.00	1.50	5.50
51	95.47	**55.75**	693.25	749.00	98.66	**31.25**	276.75	308.00	99.79	**2.75**	1.50	4.25
75	95.53	82.75	954.25	1037.00	98.59	35.75	260.00	295.75	99.74	3.00	1.50	4.50
101	94.60	75.75	1540.50	1616.25	98.55	37.25	259.00	296.25	99.76	3.50	1.75	5.25
151	95.41	111.75	1082.75	1194.50	98.54	41.50	**210.75**	**252.25**	99.78	6.25	1.50	7.75
201	**96.00**	76.50	**487.25**	**563.75**	**98.68**	53.75	260.00	313.75	**99.79**	**2.75**	**1.00**	**3.75**
251	93.71	92.75	758.00	850.75	98.43	43.50	218.00	261.50	99.62	3.50	2.25	5.75
301	95.63	76.00	777.00	853.00	98.41	42.50	218.00	260.50	99.70	**2.75**	**1.00**	**3.75**

Similar to Mackiewicz et al. (2008), Li et al. (2015), Zhao et al. (2021) and Son et al. (2022), the performance of localizing the anatomical landmarks is assessed as the frame distance (error) between the image where the boundaries of the organ were manually annotated by the experts and those predicted by the system. Mean absolute error (MAE) and median absolute error are used to quantify the performance. Since the capsule frame rate is variable, both errors (MAE and median) are also presented as the difference in time (except in the *Kvasir-Capsule* dataset, where frame time information is not available).

It is important to note that the metrics are computed per video to avoid any bias caused by the video lengths.

4.3. Implementation details

TensorFlow 2.4 was used to implement the models, which were executed on a machine with an NVIDIA GeForce RTX 2080 TI and CUDA 11.0. The training process is composed of two separate stages. Firstly, the DNN with the temporal block is trained. Then, the weights of the DNN are frozen and the CMT block is optimized.

ResNet-50 (He et al., 2016) initialized with ImageNet weights has been used as a backbone architecture for the first stage DNN. The optimization of the network was carried out with Stochastic Gradient Descent and a batch size of 256. In all the experiments, the networks were trained for $10K$ iterations. For all the datasets, the learning rate was set to 0.1 and it was decreased by a factor of 0.1 every $2K$ iterations.

All the images were resized to 128 × 128 pixels. In the case of the private datasets, a uniform circle mask was applied over each frame to eliminate the artifacts present at the borders of the images, ensuring that no specific noise or patterns could identify either a dataset or a particular video.

Data augmentation techniques were applied during the training phase to improve the robustness of the method. Specifically, rotations of 0, 90, 180, and 270 degrees, horizontal and vertical flips, and changes in the brightness of the images were used.

The CMT network from the second stage is composed of two bidirectional LSTM layers with 200 and 100 units, respectively. Finally, the dense layer has two neurons as output. The network was optimized with RMSprop as it is recommended by Zaman et al. (2021). The learning rate was fixed to 0.001 during $4K$ iterations. The batch size was set to 512. To find the optimal hyper-parameter window size, w, a search grid was done, and the chosen value was = 201 for the *Kvasir-Capsule* and *Capri* datasets and w = 151 for the *VH* dataset.

Since the presented datasets are statistically different, the hyper-parameter window size, w, must be chosen carefully for each case. The reported metrics in Table 2 are the AUC score of the model in the image classification task and the median error in the entrance, exit, and sum of both in the landmark identification task. A window size of 201 achieves the best results with the AUC metric for all the datasets. At the entrance of the organ, the smallest error is obtained with $w = 51$, whereas the lowest error in the exit is achieved with $w = 151$ in *VH* dataset and $w = 201$ in *Kvasir-Capsule* and *Capri* datasets. The same values of w are the ones that obtain the lowest error in the sum of the entrance and exit of the organs. Therefore, the hyper-parameter w chosen for *Kvasir-Capsule* and *Capri* datasets is $w = 201$ because the AUC and total median error coincide. In the case of the *VH* dataset, the chosen value is $w = 151$ since the difference between the AUC values for $w = 151$ and $w = 201$ is negligible.

Table 3
Overview of the ablation settings and the name used.

Method	Ablation settings				
	ResNet	Temp. Block	Context		
			Prob.	Motion	Time
ResNet	✓				
ResNet + C	✓		✓		
ResNet + CM	✓		✓	✓	
ResNet + CT	✓		✓		✓
ResNet + CMT	✓		✓	✓	✓
ResNet + Time	✓	✓			
ResNet + Time + C	✓	✓	✓		
ResNet + Time + CM	✓	✓	✓	✓	
ResNet + Time + CT	✓	✓	✓		✓
Proposed Method	✓	✓	✓	✓	✓

5. Results

The results section is divided into two sets of experiments. The first one presents the performance of the classification task for each one of the datasets. The second is focused on identifying the exact frames where the capsule enters and exits the studied organ. Finally, qualitative results are shown to complement the quantitative results of the proposed system compared to methods published so far.

In all the experiments, *Kvasir-Capsule*, *VH*, and *Capri* datasets are evaluated using our proposed method, which consists of a DNN concatenated with the temporal and CMT blocks. To analyze the influence of each component, an ablation study is performed by building several additional models. Finally, our method is compared with the state-of-the-art works in each task.

For the ablation study, *ResNet* is the simplest method and is considered the baseline for the comparisons. The influence of the temporal block is evaluated with the *ResNet + Time* model, which combines the image representation obtained by *ResNet* with the timestamp of the image. The probability signals generated with the outputs of each model are used to study the contribution of the context block. Let us note that *ResNet + Time + CMT* is the proposed method. Table 3 contains a summary of the ablation settings of each method.

5.1. Image classification

The first experiment evaluates the performance of the proposed method in the image classification problem. Specifically, on *Kvasir-Capsule* and *VH* datasets, all the compared models aim to identify the

Table 4
Comparison of the ablation study in the image classification problem for each dataset. Displayed results are the mean obtained after evaluating a two-fold cross-validation.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	ResNet + C	93.53 ± 5.51	87.24 ± 13.04	82.78 ± 10.18	73.05 ± 16.82	92.50 ± 11.90
	ResNet + CM	92.70 ± 6.22	88.47 ± 11.18	83.95 ± 9.71	74.41 ± 16.75	93.49 ± 10.66
	ResNet + CT	94.40 ± 6.22	87.65 ± 11.59	83.72 ± 11.06	75.45 ± 18.98	91.98 ± 11.32
	ResNet + CMT	95.47 ± 5.39	90.07 ± 7.25	85.18 ± 8.86	75.23 ± 17.76	95.12 ± 6.54
	ResNet + Time	92.40 ± 5.06	87.88 ± 6.04	81.16 ± 7.66	67.92 ± 15.42	94.40 ± 5.27
	ResNet + Time + C	94.91 ± 4.29	89.53 ± 6.73	**87.69 ± 7.31**	**82.11 ± 15.57**	93.27 ± 6.96
	ResNet + Time + CM	94.67 ± 5.24	89.87 ± 6.54	87.39 ± 7.48	80.83 ± 15.88	93.95 ± 6.30
	ResNet + Time + CT	**96.36 ± 3.98**	90.80 ± 5.80	87.62 ± 7.77	80.21 ± 16.73	95.03 ± 4.92
	Proposed Method	96.00 ± 4.57	**91.36 ± 5.75**	87.47 ± 7.49	78.91 ± 16.28	**96.03 ± 4.29**
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	ResNet + C	96.28 ± 6.93	93.44 ± 6.27	91.89 ± 8.01	87.31 ± 15.88	96.47 ± 3.97
	ResNet + CM	97.70 ± 4.05	93.78 ± 5.94	91.97 ± 7.86	87.15 ± 15.69	96.79 ± 3.93
	ResNet + CT	97.81 ± 3.69	92.59 ± 6.93	91.12 ± 8.37	86.10 ± 17.00	96.15 ± 4.68
	ResNet + CMT	97.98 ± 3.31	93.49 ± 6.52	92.02 ± 8.06	87.55 ± 15.86	96.49 ± 4.19
	ResNet + Time	95.97 ± 6.28	88.64 ± 8.16	89.24 ± 7.57	**89.94 ± 12.64**	88.56 ± 9.00
	ResNet + Time + C	97.17 ± 6.35	94.63 ± 5.84	**92.58 ± 8.06**	88.00 ± 16.06	97.16 ± 3.85
	ResNet + Time + CM	98.13 ± 3.49	94.55 ± 5.84	92.41 ± 8.07	87.56 ± 15.96	97.26 ± 3.81
	ResNet + Time + CT	98.20 ± 3.79	**94.69 ± 5.79**	92.55 ± 7.94	87.25 ± 15.89	**97.84 ± 3.28**
	Proposed Method	**98.54 ± 2.36**	94.58 ± 5.17	92.26 ± 7.74	87.25 ± 15.78	97.27 ± 3.42
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	ResNet + C	99.83 ± 0.81	98.63 ± 3.40	98.51 ± 3.30	98.50 ± 3.63	98.52 ± 5.59
	ResNet + CM	99.82 ± 0.74	98.70 ± 3.21	98.37 ± 3.56	97.83 ± 5.01	98.90 ± 4.94
	ResNet + CT	99.79 ± 0.92	98.54 ± 3.51	98.51 ± 3.13	98.54 ± 3.03	98.47 ± 5.69
	ResNet + CMT	99.86 ± 0.54	98.73 ± 3.18	98.42 ± 3.54	98.01 ± 4.90	98.84 ± 5.07
	ResNet + Time	99.66 ± 0.76	97.18 ± 3.50	96.51 ± 3.37	93.82 ± 6.46	99.20 ± 2.28
	ResNet + Time + C	99.88 ± 0.54	98.97 ± 2.22	98.79 ± 2.36	98.08 ± 4.19	99.51 ± 2.28
	ResNet + Time + CM	99.59 ± 1.52	98.91 ± 2.55	98.76 ± 2.51	97.92 ± 4.66	**99.60 ± 2.11**
	ResNet + Time + CT	**99.90 ± 0.47**	99.02 ± 2.01	98.90 ± 2.16	98.43 ± 3.49	99.36 ± 2.81
	Proposed Method	99.79 ± 0.82	**99.07 ± 2.12**	**98.96 ± 2.21**	**98.58 ± 3.48**	99.35 ± 2.97

Table 5
Comparison of the different methods of the state-of-the-art with our model in the image classification problem for each dataset. Displayed results are the mean obtained after evaluating a two-fold cross-validation.

Dataset	Methods	AUC (%)	ACC (%)	MACC (%)	SPEC (%)	SENS (%)
Kvasir-Capsule	ResNet	91.48 ± 4.96	87.13 ± 7.00	82.10 ± 7.78	71.75 ± 15.43	92.45 ± 7.22
	Zou et al. (2015)	75.37 ± 9.42	69.51 ± 11.67	69.11 ± 8.68	70.19 ± 16.35	68.03 ± 14.60
	Chen et al. (2017)	83.65 ± 10.37	82.38 ± 9.20	76.90 ± 10.28	67.95 ± 16.18	85.84 ± 10.96
	Zhao et al. (2021)	94.05 ± 4.50	89.46 ± 7.72	85.09 ± 7.87	76.29 ± 14.40	93.89 ± 7.46
	Son et al. (2022)	95.75 ± 4.85	90.96 ± 6.56	81.03 ± 12.55	64.42 ± 26.40	**97.64 ± 4.40**
	Proposed Method	**96.00 ± 4.57**	**91.36 ± 5.75**	**87.47 ± 7.49**	**78.91 ± 16.28**	96.03 ± 4.29
VH	ResNet	94.42 ± 6.70	84.60 ± 9.59	86.26 ± 8.36	88.26 ± 13.96	84.25 ± 10.27
	Zou et al. (2015)	90.05 ± 9.91	84.56 ± 11.00	74.78 ± 12.22	56.87 ± 24.96	92.68 ± 9.98
	Chen et al. (2017)	95.86 ± 5.46	90.29 ± 8.12	87.69 ± 8.28	82.53 ± 15.62	92.84 ± 10.02
	Zhao et al. (2021)	97.81 ± 4.24	93.56 ± 7.12	91.95 ± 8.04	87.54 ± 14.89	96.37 ± 5.07
	Son et al. (2022)	96.46 ± 6.65	89.27 ± 9.35	90.46 ± 8.73	**91.05 ± 14.59**	89.88 ± 9.21
	Proposed Method	**98.54 ± 2.36**	**94.58 ± 5.17**	**92.26 ± 7.74**	87.25 ± 15.78	**97.27 ± 3.42**
Capri	ResNet	99.09 ± 1.41	95.71 ± 3.67	92.36 ± 4.62	85.70 ± 9.06	99.00 ± 3.16
	Zou et al. (2015)	86.06 ± 7.93	80.64 ± 12.24	65.93 ± 6.29	33.50 ± 12.51	98.35 ± 2.02
	Chen et al. (2017)	95.28 ± 4.37	88.42 ± 7.84	88.69 ± 6.49	88.31 ± 10.25	89.07 ± 9.62
	Zhao et al. (2021)	99.85 ± 0.47	98.59 ± 2.23	98.17 ± 2.94	97.76 ± 3.74	98.58 ± 4.14
	Son et al. (2022)	**99.93 ± 0.21**	97.94 ± 2.74	96.06 ± 4.30	92.57 ± 8.22	**99.57 ± 2.58**
	Proposed Method	99.79 ± 0.82	**99.07 ± 2.12**	**98.96 ± 2.21**	**98.58 ± 3.48**	99.35 ± 2.97

small bowel frames, whereas, on the *Capri* dataset, they aim to classify the large intestine images.

The obtained results in the ablation study are presented in Table 4. In all the datasets, the temporal block enhances the performance of the methods with respect to the baseline *ResNet*. Similarly, it can be seen that the obtained results by the models with the context block are higher than the baselines (*ResNet* and *ResNet* + *Time*). In general, when time or motion is added to the context block, the models achieve better results. This means that the combination of visual, temporal, and contextual information produces a powerful discriminative model. It can also be observed that the higher performance obtained in our model is an AUC value of 99.79% on the *Capri* dataset. On *Kvasir-Capsule* and *VH*, the obtained scores are 96.00% and 98.54%, respectively. Several reasons can justify the difference in performance among datasets, being the main differences between them: (1) the organ of study (colon on *Capri* vs. small bowel on *Kvasir-Capsule* and *VH*); (2) capsule device (Olympus EC-S10, Medtronic PillCam SB3, and Medtronic PillCam Colon2); and (3) amount of intestinal content. Therefore, capsule characteristics like optic, illumination, and resolution are not equivalent neither the intestinal mucosa and content. In addition, the statistics from each dataset are different as reported in Table 1. Despite all the mentioned differences, the results are coherent among the various datasets.

As previously stated, *Kvasir-Capsule* dataset lacks temporal information. To address this limitation, the frame index has been used as a

Table 6
Comparison of the ablation study in the anatomical landmarks identification task for each dataset. MAE and median error are represented as the difference in frames and time (hh:mm:ss).

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	ResNet	668.34 ± 1091.84	–	111.75 ± 28.25	–	1875.96 ± 2747.54	–	1124.00 ± 493.00	–
	ResNet + C	1505.83 ± 3316.56	–	207.50 ± 153.00	–	2147.42 ± 2967.07	–	908.50 ± 757.50	–
	ResNet + CM	1504.54 ± 3272.05	–	217.25 ± 131.75	–	1877.62 ± 2736.03	–	928.00 ± 684.50	–
	ResNet + CT	1677.79 ± 3509.05	–	139.00 ± 51.50	–	1902.29 ± 2683.68	–	1220.50 ± 391.50	–
	ResNet + CMT	830.08 ± 1341.51	–	127.50 ± 31.00	–	1770.79 ± 2771.27	–	743.75 ± 580.25	–
	ResNet + Time	785.88 ± 1182.88	–	93.00 ± 15.50	–	2002.38 ± 2959.43	–	1077.00 ± 539.00	–
	ResNet + Time + C	535.50 ± 1089.54	–	**26.75 ± 2.25**	–	1710.71 ± 2769.83	–	559.00 ± 169.00	–
	ResNet + Time + CM	606.08 ± 1104.42	–	61.25 ± 9.25	–	1727.67 ± 2767.06	–	651.50 ± 383.00	–
	ResNet + Time + CT	556.38 ± 951.79	–	116.00 ± 43.00	–	1730.50 ± 2756.90	–	663.25 ± 83.25	–
	Proposed Method	**465.88 ± 918.13**	–	76.50 ± 46.50	–	**1679.67 ± 2775.72**	–	**487.25 ± 163.75**	–
VH	ResNet	667.70 ± 1070.13	00 : 04 : 00	220.75 ± 191.75	00 : 01 : 45	2710.98 ± 4505.58	00 : 22 : 02	1235.75 ± 1126.75	00 : 12 : 20
	ResNet + C	559.98 ± 1007.92	00 : 03 : 11	78.50 ± 17.00	00 : 00 : 34	1290.33 ± 2117.18	00 : 14 : 52	198.50 ± 82.00	00 : 04 : 02
	ResNet + CM	512.38 ± 870.20	00 : 02 : 59	91.75 ± 48.75	00 : 00 : 39	1028.02 ± 1704.22	00 : 12 : 14	199.50 ± 85.50	00 : 03 : 24
	ResNet + CT	557.83 ± 1181.40	00 : 03 : 10	78.25 ± 16.75	00 : 00 : 32	1279.73 ± 1774.48	00 : 14 : 55	369.75 ± 3.75	00 : 05 : 53
	ResNet + CMT	500.90 ± 1166.62	00 : 02 : 45	50.50 ± 17.0	00 : 00 : 25	1077.35 ± 1605.62	00 : 12 : 05	259.50 ± 109.00	00 : 05 : 40
	ResNet + Time	731.71 ± 1451.30	00 : 03 : 57	103.50 ± 59.50	00 : 01 : 14	2050.94 ± 4142.68	00 : 17 : 01	397.00 ± 350.00	00 : 06 : 26
	ResNet + Time + C	502.62 ± 877.55	00 : 02 : 47	59.50 ± 4.00	00 : 00 : 25	911.23 ± 1619.69	00 : 11 : 38	167.25 ± 106.25	00 : 03 : 03
	ResNet + Time + CM	**417.79 ± 814.67**	**00 : 02 : 18**	44.00 ± 7.00	**00 : 00 : 10**	1110.48 ± 1992.44	00 : 13 : 02	**166.50 ± 103.00**	**00 : 03 : 00**
	ResNet + Time + CT	443.46 ± 1092.63	00 : 02 : 31	50.25 ± 8.25	00 : 00 : 23	1051.17 ± 1933.66	00 : 11 : 37	225.25 ± 170.25	00 : 03 : 29
	Proposed Method	443.69 ± 1064.05	00 : 02 : 38	**41.50 ± 11.00**	00 : 00 : 15	**837.77 ± 1485.79**	**00 : 09 : 46**	210.75 ± 164.75	00 : 03 : 14
Capri	ResNet	53.70 ± 110.44	00 : 01 : 19	14.50 ± 4.00	00 : 00 : 06	13.80 ± 48.35	00 : 02 : 19	1.00 ± 0.00	**00 : 00 : 01**
	ResNet + C	28.94 ± 85.33	00 : 00 : 59	4.50 ± 0.50	00 : 00 : 02	41.76 ± 215.63	00 : 04 : 44	3.00 ± 0.00	**00 : 00 : 01**
	ResNet + CM	32.43 ± 97.72	00 : 01 : 00	2.75 ± 0.25	**00 : 00 : 01**	38.58 ± 210.08	00 : 03 : 01	1.75 ± 0.25	**00 : 00 : 01**
	ResNet + CT	32.62 ± 91.01	00 : 01 : 02	5.00 ± 1.00	00 : 00 : 02	48.45 ± 240.70	00 : 07 : 08	2.00 ± 0.50	00 : 00 : 02
	ResNet + CMT	32.35 ± 100.58	00 : 01 : 01	2.50 ± 0.50	**00 : 00 : 01**	37.71 ± 209.38	00 : 03 : 06	2.00 ± 0.00	**00 : 00 : 01**
	ResNet + Time	38.40 ± 86.46	00 : 01 : 11	5.00 ± 2.00	00 : 00 : 02	19.80 ± 114.86	00 : 05 : 37	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + C	29.57 ± 81.18	00 : 00 : 51	4.50 ± 1.50	**00 : 00 : 01**	8.89 ± 38.02	00 : 02 : 13	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CM	**23.58 ± 69.27**	**00 : 00 : 41**	3.50 ± 0.50	**00 : 00 : 01**	8.26 ± 38.02	00 : 02 : 04	**1.00 ± 0.00**	**00 : 00 : 01**
	ResNet + Time + CT	30.40 ± 79.19	00 : 00 : 58	5.25 ± 3.25	**00 : 00 : 01**	8.88 ± 38.15	00 : 02 : 09	**1.00 ± 0.00**	**00 : 00 : 01**
	Proposed Method	29.40 ± 83.94	00 : 00 : 55	**2.75 ± 0.25**	**00 : 00 : 01**	**7.91 ± 37.82**	**00 : 01 : 47**	**1.00 ± 0.00**	**00 : 00 : 01**

Table 7
Comparison of the different methods of the state-of-the-art with our model in the anatomical landmarks identification task for each dataset. MAE and median error are represented as the difference in frames and time (hh:mm:ss).

Dataset	Methods	Entrance				Exit			
		MAE		Median		MAE		Median	
		Frame	Time	Frame	Time	Frame	Time	Frame	Time
Kvasir-Capsule	Zhao et al. (2021)	2644.16 ± 4637.65	–	1251.00 ± 115.25	–	4603.58 ± 1545.95	–	1669.00 ± 185.26	–
	Son et al. (2022)	2711.00 ± 3435.83	–	1786.00 ± 231.93	–	2409.00 ± 3106.30	–	1506.75 ± 1161.42	–
	Proposed Method	**465.88 ± 918.13**	–	**76.50 ± 46.50**	–	**1679.67 ± 2775.72**	–	**487.25 ± 163.75**	–
VH	Zhao et al. (2021)	1304.23 ± 1394.26	00 : 08 : 20	915.25 ± 22.98	00 : 04 : 34	3308.58 ± 583.28	00 : 31 : 44	1765.25 ± 461.03	00 : 16 : 59
	Son et al. (2022)	1390.47 ± 3487.30	00 : 07 : 12	304.75 ± 220.97	00 : 02 : 14	1552.00 ± 520.78	00 : 16 : 47	627.75 ± 1469.01	00 : 08 : 09
	Proposed Method	**443.69 ± 1064.05**	**00 : 02 : 38**	**41.50 ± 11.00**	**00 : 00 : 15**	**837.77 ± 1485.79**	**00 : 09 : 46**	**210.75 ± 164.75**	**00 : 03 : 14**
Capri	Zhao et al. (2021)	214.24 ± 437.6	00 : 07 : 28	85.0 ± 23.33	00 : 01 : 11	524.94 ± 1258.34	00 : 35 : 23	23.5 ± 2.12	00 : 01 : 02
	Son et al. (2022)	56.39 ± 161.94	00 : 04 : 00	14.50 ± 0.70	00 : 00 : 08	32.25 ± 111.27	00 : 07 : 29	7.50 ± 0.70	00 : 00 : 43
	Proposed Method	**29.40 ± 83.94**	**00 : 00 : 55**	**2.75 ± 0.25**	**00 : 00 : 01**	**7.91 ± 37.82**	**00 : 01 : 47**	**1.00 ± 0.00**	**00 : 00 : 01**

substitute for temporal information. Despite this adjustment, similar effects on the system's performance have been observed in this dataset. This can be attributed to the fact that the order of the frames is a reliable proxy for timestamps.

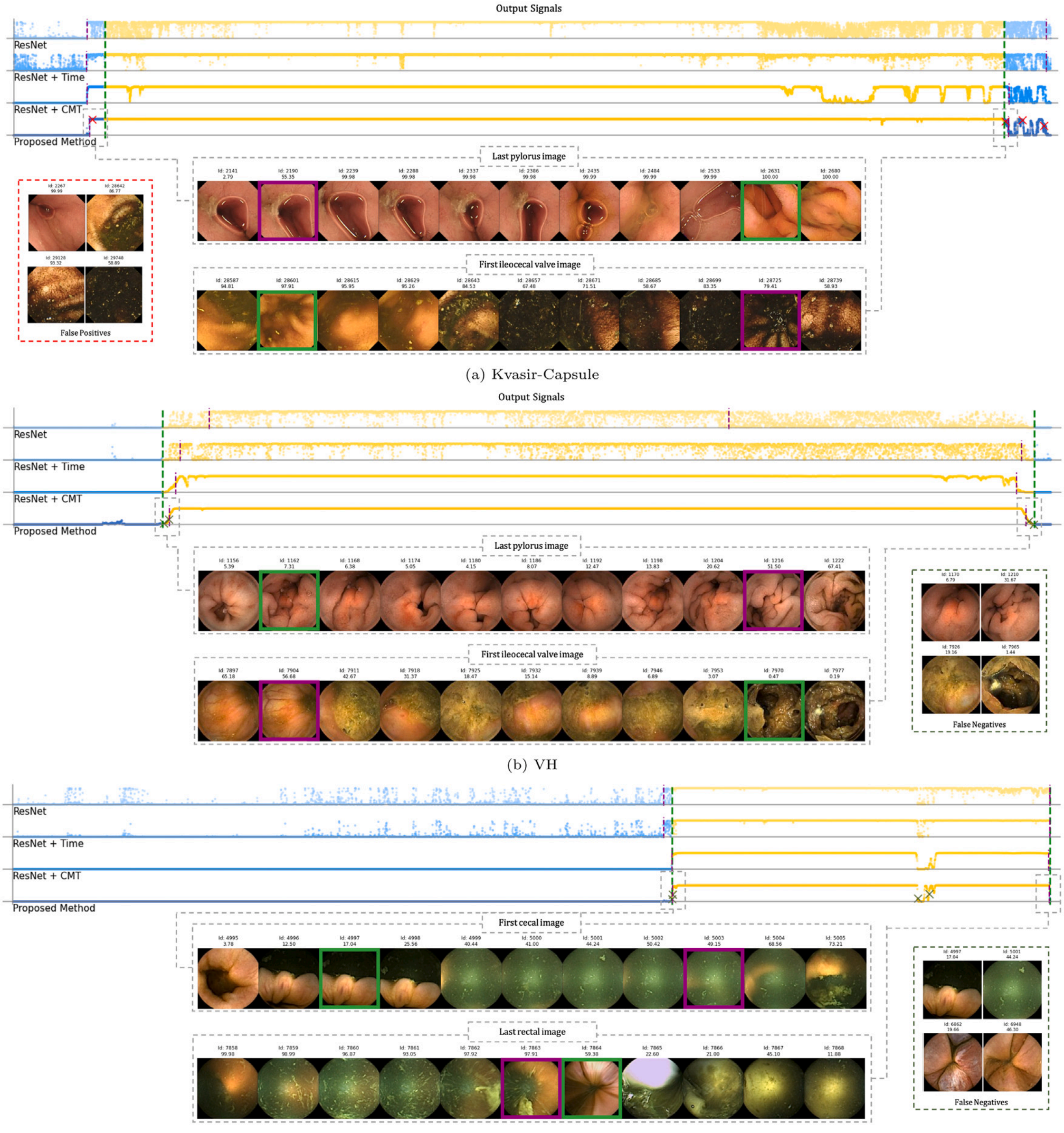
The proposed method is compared with the following state-of-the-art methods: Zou et al. (2015), Chen et al. (2017), Zhao et al. (2021), and Son et al. (2022). All these methods have been implemented, trained, and evaluated using the same datasets and evaluation methodology. The results reported in Table 5 show that the proposed method outperforms all others in all datasets.

5.2. Anatomical landmarks identification

In the second experiment, the difference between the predicted landmarks and the annotations provided by the experts is analyzed. On *Kvasir-Capsule* and *VH* datasets, the pylorus and the ileocecal valve, which delimit the small bowel, are identified. On the other hand, on *Capri* dataset, the boundaries of the colon, first cecal and last rectal images are used.

The results from the ablation study come from minimizing the rectangular pulse function over the output signal of each setting. Table 6 contains the MAE and median error of each one in frames and time. The reported results show that the use of the temporal and context block reduces the error of the baseline *ResNet*. Particularly, in the small bowel datasets, *Kvasir-Capsule* and *VH*, there is a large difference between MAE and median error. This suggests that there are several outliers. Despite them, the proposed method achieves promising results in all the cases.

The proposed method is compared with Zhao et al. (2021) and Son et al. (2022), as shown in Table 7. For this experiment, the proposed



(a) Kvasir-Capsule

(b) VH

(c) Capri

Fig. 6. Visual representation of the system outputs of three WCE videos from: (a) *Kvasir-Capsule* (b) *VH* and (c) *Capri* datasets. Each subfigure contains the output signals and the identification of the anatomical landmarks for the evaluated methods. Yellow points represent frames from the organ of interest (small bowel or large intestine), whereas the blues ones are outside these areas. The second task is displayed over the outputs signals as dashed lines. The predicted landmarks are ticked in purple, while the ground truth is in green. Below the output signals are displayed a uniform sampling of frames around the landmarks, achieving sequences of 11 items. The frame identification (id) and the probability of belonging to the organ of interest are shown above each image. The frames of the labeled and predicted landmarks are surrounded by a green and purple box, respectively. Finally, several misclassified frames are shown, which are localized in the output signal of the *Proposed Method* as crosses in red for false positives and dark green for false negatives samples. The figure is best viewed on the computer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
Comparison of the different strategies for identifying the anatomical landmarks. The proposed strategy is applied to state-of-art methods. The displayed results are the median error obtained after evaluating a two-fold cross-validation.

Method	Dataset					
	Kvasir-Capsule		VH		Capri	
	Entrance	Exit	Entrance	Exit	Entrance	Exit
Zhao et al. (2021)	1251.00	1669.00	915.25	1765.25	85.00	23.50
Zhao et al. (2021) + Step 3	93.00	702.50	65.00	478.00	3.50	1.00
Son et al. (2022)	1786.00	1506.00	304.75	627.75	14.50	7.50
Son et al. (2022) + Step 3	686.25	1683.25	214.50	1236.00	35.50	6.00
Proposed Method	**76.50**	**487.25**	**41.50**	**210.75**	**2.75**	**1.00**

methods by Zou et al. (2015) and Chen et al. (2017) have not been considered since they do not identify the boundaries of the organs. The results in Table 7 show that the proposed method outperformed all methods in all the cases.

It can be observed in Tables 6 and 7 that in both small bowel datasets, that the identification of the entrance of the organ is more accurate than the exit. On the other hand, for the colon dataset (*Capri*) the identification of last rectal image is more accurate than the entrance.

Finally, the impact of the landmarks identification strategy is analyzed. To evaluate it, the third step of our model has been applied to the outputs of the models from Zhao et al. (2021) and Son et al. (2022). Table 8 summarizes the results obtained, which indicate that by solving the minimization problem the errors decrease. These findings suggest that the proposed strategy of locating anatomical landmarks is more effective than those currently published.

The results from both tasks, image classification (Table 5) and anatomical landmark identification (Table 7) show a strong correlation. In other words, the better the classifier, the more accurate the limit identification of the organs.

5.3. Qualitative results

This section aims to gain additional information about the performance of the proposed method and to visualize the types of errors in a qualitative manner (Fig. 6). These results are presented using one test video per dataset (*Kvasir-Capsule* dataset in Fig. 6(a), *VH* dataset in Fig. 6(b) and *Capri* dataset in Fig. 6(c)). The selected ones have metrics near the median values reported in Table 6, thus avoiding outliers and championship cases. The output signals represent the classification task, where the probability for each image to belong to the studied organ is plotted. Yellow dots are the frames corresponding to the organs (small bowel or large intestine), whereas the blue ones are considered out of this range. The visualization allows assessing the performance of our method by removing certain blocks and understanding the contribution of each component to the overall system. For that purpose, each figure shows four of the methods previously introduced: *ResNet*, *ResNet + Time*, *ResNet + CMT* and *Proposed Method*.

For the *Kvasir-Capsule* dataset (Fig. 6(a)), the *ResNet + Time* method infers worse probabilities than the *ResNet* model. When the temporal information is combined with the contextual block in our method, the obtained output signal is smoother and more similar to the ground truth. However, there are still some misclassified sequences outside the small intestine. In the case of the videos from the other datasets (Fig. 6(b) and Fig. 6(c)) the application of the contextual information further evidence the improved performance of the proposed method even more.

Moreover, each subfigure contains a set of false positive and false negatives samples determined by our method. The small bowel videos show several misclassified examples, where the mucosa is completely hidden, thus preventing the system from making a correct prediction.

The identification of the anatomical landmarks is a complex task since only one correct frame in the video has to be determined as the entrance or exit of the organ. In Fig. 6, the green dashed lines over the output signals indicate the frame labeled by the experts while the purple ones correspond to the system predictions. The rectangular pulse function fits the output signal and correctly identifies the landmarks, but it fails when the prediction of belonging to the organ is wrong. In the identified last pylorus image, it can be observed that despite the distance between the predicted and the real landmarks, the frames are visually similar (Figs. 6(a) and 6(b)). But when the mucosa is hidden by the noise content of the GI tract, as happens in the exit sequences in Figs. 6(a) and 6(b), the error is higher. Therefore, the complexity of the problem increases. However, in *Capri* dataset (Fig. 6(c)), the exit of the large intestine is easier to identify due to the drastic change in the visual features caused by the evacuation of the capsule from the body or because the video stopped.

6. Discussion and conclusion

In this paper, an effective deep learning system for WCE is proposed. The method is designed to, first, infer the probability for every image to belong to the area of interest, taking advantage of temporal, neighboring, and motion information. Secondly, the landmarks are predicted by solving a minimization problem. Experimental results have been reported over three datasets, one public and two private. In all of them, the proposed method improves the results of the baseline system.

The results obtained in the two datasets of the small bowel, *Kvasir-Capsule* and *VH*, show a high performance in the classification problem. Moreover, our method in the *VH* dataset displays even better results, increasing at least three points in each metric. Several reasons can justify the difference in performance achieved on each dataset such as dataset size, amount of intestinal content, or device used to collect the video.

After performing all the experiments and analyzing the results, we believe that our method is a good candidate for the automatic classification of organs regardless of the device used. Although our method of landmark identification does not achieve the best performance for all the datasets, it exhibits promising results.

One limitation of the proposed system is that it has been designed to deal with only one organ, given the lack of multi-organ labels in the used datasets. However, this limitation could be addressed by incorporating anatomical landmarks for multiple organs and adapting the CMT block accordingly. Additionally, the results strongly depend on the organ, the used WCE device, and the dataset size. To overcome this issue, future research could focus on exploring a general method for multiple devices and organs.

Furthermore, future work could also investigate the detection of multiple organs and their anatomical landmarks. Additionally, it may be possible to localize distinct landmarks within a single organ, such as the flexures of the large intestine.

CRediT authorship contribution statement

Pablo Laiz: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jordi Vitrià:** Conceptualization, Writing – review & editing. **Pere Gilabert:** Conceptualization, Writing – review & editing. **Hagen Wenzek:** Data curation, Writing – review & editing. **Carolina Malagelada:** Data curation, Validation, Resources, Writing – review & editing. **Angus J.M. Watson:** Data curation. **Santi Seguí:** Conceptualization, Investigation, Writing – original draft, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Hagen Wenzek reports a relationship with CorporateHealth International ApS that includes: board membership.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work has been partially funded by MINECO, Spain Grant RTI2018-095232-B-C21, SGR 1742 and 01104 (Generalitat de Catalunya, Spain), Innovate UK project 104633, AI AWARD02440 project from NIHR, UK and NHSX, UK and the FPU, Spain grants FPU17/02727 and FPU20/01090 (Ministerio de Universidades, Spain). The authors would also like to thank NVIDIA for their GPU donations.

References

Adewole, S., Yeghyayan, M., Hyatt, D., Ehsan, L., Jablonski, J., Copland, A., Syed, S., Brown, D., 2020. Deep learning methods for anatomical landmark detection in video capsule endoscopy images. In: Proceedings of the Future Technologies Conference. Springer, pp. 426–434.

Berens, J., Mackiewicz, M., Bell, D., 2005. Stomach, intestine, and colon tissue discriminators for wireless capsule endoscopy images. In: Fitzpatrick, J.M., Reinhardt, J.M. (Eds.), Medical Imaging 2005: Image Processing, Vol. 5747. SPIE, International Society for Optics and Photonics, pp. 283–290. http://dx.doi.org/10.1117/12.594799.

Chen, H., Wu, X., Tao, G., Peng, Q., 2017. Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos. Neurocomputing 229, 77–87. http://dx.doi.org/10.1016/j.neucom.2016.06.077, URL https://www.sciencedirect.com/science/article/pii/S0925231216313728, Advances in computing techniques for big medical image data.

Darrow, J., 2014. Capsule endoscopy instead of colonoscopy? The FDA approves the PillCam COLON. https://blog.petrieflom.law.harvard.edu/2014/03/04/capsule-endoscopy-instead-of-colonoscopy-the-fda-approves-the-pillcam-colon/.

Dokoutsidou, H., Karagiannis, S., Giannakoulopoulou, E., Galanis, P., Kyriakos, N., Liatsos, C., Faiss, S., Mavrogiannis, C., 2011. A study comparing an endoscopy nurse and an endoscopy physician in capsule endoscopy interpretation. Eur. J. Gastroenterol. Hepatol. 23 (2), 166–170.

Haji-Maghsoudi, O., Talebpour, A., Soltanian-Zadeh, H., Haji-Maghsoodi, N., 2012. Automatic organs' detection in WCE. In: AISP 2012 - 16th CSI International Symposium on Artificial Intelligence and Signal Processing. IEEE, pp. 116–121. http://dx.doi.org/10.1109/AISP.2012.6313729.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Iakovidis, D.K., Koulaouzidis, A., 2015. Software for enhanced video capsule endoscopy: challenges for essential progress. Nature Rev. Gastroenterol. Hepatol. 12 (3), 172–186.

Iddan, G., Meron, G., Glukhovsky, A., Swain, P., 2000. Wireless capsule endoscopy. Nature 405 (6785), 417.

Koulaouzidis, A., Dabos, K., Philipper, M., Toth, E., Keuchel, M., 2021. How should we do colon capsule endoscopy reading: A practical guide. Therapeutic Adv. Gastrointest. Endosc. 14, 26317745211001983.

Lee, J., Oh, J., Shah, S.K., Yuan, X., Tang, S.J., 2007. Automatic classification of digestive organs in wireless capsule endoscopy videos. In: Proceedings of the 2007 ACM Symposium on Applied Computing. SAC '07, Association for Computing Machinery, New York, NY, USA, pp. 1041–1045. http://dx.doi.org/10.1145/1244002.1244230, URL https://doi-org.sire.ub.edu/10.1145/1244002.1244230.

Li, B., Xu, G., Zhou, R., Wang, T., 2015. Computer aided wireless capsule endoscopy video segmentation. Med. Phys. 42 (2), 645–652. http://dx.doi.org/10.1118/1.4905164.

Mackiewicz, M., Berens, J., Fisher, M., 2008. Wireless capsule endoscopy color video segmentation. IEEE Trans. Med. Imaging 27 (12), 1769–1781. http://dx.doi.org/10.1109/TMI.2008.926061.

Maieron, A., Hubner, D., Blaha, B., Deutsch, C., Schickmair, T., Ziachehabi, A., Kerstan, E., Knoflach, P., Schoefl, R., 2004. Multicenter retrospective evaluation of capsule endoscopy in clinical routine. Endoscopy 36 (10), 864–868.

Pascual, G., Laiz, P., García, A., Wenzek, H., Vitrià, J., Seguí, S., 2022. Time-based self-supervised learning for wireless capsule endoscopy. http://dx.doi.org/10.48550/ARXIV.2204.09773, URL https://arxiv.org/abs/2204.09773.

Rondonotti, E., Pennazio, M., Toth, E., Koulaouzidis, A., 2020. How to read small bowel capsule endoscopy: A practical guide for everyday use. Endosc. Int. Open 8 (10), E1220–E1224.

Smedsrud, P.H., Thambawita, V., Hicks, S.A., Gjestang, H., Nedrejord, O.O., Næss, E., Borgli, H., Jha, D., Berstad, T.J.D., Eskeland, S.L., Lux, M., Espeland, H., Petlund, A., Nguyen, D.T.D., Garcia-Ceja, E., Johansen, D., Schmidt, P.T., Toth, E., Hammer, H.L., de Lange, T., Riegler, M.A., Halvorsen, P., 2021. Kvasir-Capsule, A video capsule endoscopy dataset. Sci. Data 8 (1), 1–10. http://dx.doi.org/10.1038/s41597-021-00920-z.

Son, G., Eo, T., An, J., Oh, D.J., Shin, Y., Rha, H., Kim, Y.J., Lim, Y.J., Hwang, D., 2022. Small bowel detection for wireless capsule endoscopy using convolutional neural networks with temporal filtering. Diagnostics (Basel, Switzerland) 12 (8), 1858. http://dx.doi.org/10.3390/diagnostics12081858, URL https://europepmc.org/articles/PMC9406835.

Trasolini, R., Byrne, M.F., 2021. Artificial intelligence and deep learning for small bowel capsule endoscopy. Digestive Endosc. 33 (2).

Yung, D.E., Rondonotti, E., Koulaouzidis, A., 2016. Capsule colonoscopy—A concise clinical overview of current status. Ann. Transl. Med. 4 (20).

Zaman, S.M., Hasan, M.M., Sakline, R.I., Das, D., Alam, M.A., 2021. A comparative analysis of optimizers in recurrent neural networks for text classification. In: 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering. CSDE, pp. 1–6. http://dx.doi.org/10.1109/CSDE53843.2021.9718394.

Zhao, X., Fang, C., Gao, F., Fan, D.-j., Lin, X., Li, G., 2021. Deep Transformers for fast small intestine grounding in Capsule Endoscope Video. VIDEO School of Data and Computer Science , Sun Yat-Sen University , Guangzhou , China School of Artifical Intelligence , Xidian University , Xi ' an , China The Sixth Affiliated Ho. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 150–154.

Zou, Y., Li, L., Wang, Y., Yu, J., Li, Y., Deng, W.J., 2015. Classifying digestive organs in wireless capsule endoscopy images based on deep convolutional neural network. In: International Conference on Digital Signal Processing, DSP, Vol. 2015-September. pp. 1274–1278. http://dx.doi.org/10.1109/ICDSP.2015.7252086.