Computers in Biology and Medicine 146 (2022) 105631

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Time-based self-supervised learning for Wireless Capsule Endoscopy

Guillem Pascual [a,*], Pablo Laiz [a], Albert García [a], Hagen Wenzek [b], Jordi Vitrià [a], Santi Seguí [a]

[a] Departament de Matemàtiques i Informàtica, Universitat de Barcelona (UB), Gran Via Corts Catalanes, 585, 08007 Barcelona, Spain
[b] CorporateHealth International ApS, Denmark

ARTICLE INFO

Keywords:
capsule endoscopy
deep learning
self-supervised learning
semi-supervised learning

ABSTRACT

State-of-the-art machine learning models, and especially deep learning ones, are significantly data-hungry; they require vast amounts of manually labeled samples to function correctly. However, in most medical imaging fields, obtaining said data can be challenging. Not only the volume of data is a problem, but also the imbalances within its classes; it is common to have many more images of healthy patients than of those with pathology. Computer-aided diagnostic systems suffer from these issues, usually over-designing their models to perform accurately. This work proposes using self-supervised learning for wireless endoscopy videos by introducing a custom-tailored method that does not initially need labels or appropriate balance. We prove that using the inferred inherent structure learned by our method, extracted from the temporal axis, improves the detection rate on several domain-specific applications even under severe imbalance. State-of-the-art results are achieved in polyp detection, with 95.00 ± 2.09% Area Under the Curve, and 92.77 ± 1.20% accuracy in the CAD-CAP dataset.

1. Introduction

Obtaining gastrointestinal (GI) images has traditionally been an intrusive intervention until the advent of Wireless Capsule Endoscopy (WCE) technology [1]. WCE imaging eases the process of securing a continuous stream of images, but at the same time, it introduces its own set of problems.

The videos recorded by the capsule, although usually with a low frame rate, can have a duration of up to 12 h [2]. Unlike traditional methods, it is not a targeted exploration but rather a complete recording as the capsule travels through the entire system. A physician must go over the full length of the video, possibly at multi-image speeds, while looking for any abnormality. Not only do they have to invest considerably more time, but the fatigue and repetitiveness of the task could affect their ability to detect such abnormalities.

Providing a reliable and accurate computer-aided diagnosis (CADx) system capable of selecting the most promising frames would ease the pressure on those professionals, cutting down the time spent on the task while obtaining comparable—if not better—results.

Also of great importance, especially when designing automated systems that rely on images obtained from patients, is to examine the properties of the data. In day-to-day examinations, not all patients have an associated pathology, and the data used in research to train CADx models directly reflects it. In polyp detection, for example, the majority of videos have no polyp present in a video at all. One must also consider that, even in the case that there might be polyps, they would appear only in a small fraction of the frames [3]. A polyp might appear in several subsequent frames, perhaps slightly displaced or rotated, but the overall number would be negligible when considering the whole duration of the video.

Combining the difficulty of obtaining said datasets with the amount and distribution of the data itself makes creating accurate and production-ready CADx systems a difficult task. Data is fairly scarce compared to other problems studied in deep learning, and the classes, such as polyp, or non-polyp, suffer significant imbalances. Not to mention that supervised algorithms, which dominate the field, require that all those videos are accurately labeled to function.

Creating better models for the medical field which, ultimately, could be used in CADx, requires sorting out these issues. Techniques like data augmentation and regularization have been used to cope with overfitting and under-generalizing models, but they are hard to train and can obtain sub-par results. As such, it is the aim of this work to produce a method that enables obtaining better WCE models without over-relying on these two approaches. The main motivation being that such models would help reduce the workload that physicians are facing when examining WCE videos while, perhaps even more importantly, not

* Corresponding author.
E-mail address: guillem.pascual@ub.edu (G. Pascual).

https://doi.org/10.1016/j.compbiomed.2022.105631
Received 15 December 2021; Received in revised form 17 April 2022; Accepted 17 April 2022
Available online 24 May 2022

sacrificing any accuracy, be it detecting polyps, bleeding, or any other critical condition.

Thereafter, this work proposes the application of self-supervised learning (SSL) on WCE videos to obtain a better representation of the data, enabling future models to perform better in their classification tasks. Self-supervision has been canonically considered a variant of supervised learning [4], as the network learns from supervisory signals obtained from the data itself, often leveraging the underlying structure in the data. Based on this definition, we derive a novel pseudo-labeling method for WCE that works with several unlabeled videos, enabling the use of SSL, which helps train models for downstream tasks.

In SSL, instead of directly training a model with a set objective in mind, the process is divided into two steps. SSL is done during an initial phase named *pretrain*, where a deep neural network is trained to learn a better representation, or embedding, of the data. It encodes the most essential information into a smaller vector by using the data without their final labels, learning its inherent structure. This information is learned accordingly to the data's nature, the model's architecture, and the task used for SSL. Then, during a second pass, the *finetune* process, the embedding is used in conjunction with the labels to perform supervised classification.

With the present work, summarized in Fig. 1, we aim to use self-supervision to provide more accurate models for domain-specific tasks derived from WCE images. In particular, given unlabeled WCE videos, we exploit their temporal nature to perform SSL and then train several supervised models. These models can then be used for CADx, which would improve the results with respect to current methods, reducing the workload for physicians.

The paper is organized as follows. First, we give an overview of the related work in the field, followed by a description of our methodology, presenting the self-supervised training, supervised training, and system architecture. Further, we explain the experimental setup and results and, finally, present the main conclusions and give directions for future work.

2. Related work

2.1. Wireless Capsule Endoscopy research

WCE, due to the nature of its long data streams, has been a popular candidate for computer-aided automation. For instance, bleeding detection was first done by means of superpixels in conjunction with a support vector machine [5], using super-vector machines (SVM) and manually found color invariants [6], through saliency maps [7], and using hand-crafted textures and multiple machine learning algorithms like classification trees, random forests, and logistic model trees [8]. Other tasks explored are polyp detection through image subdivision and SVM [9], ulcer detection with texture and color invariants [10], and motility events with pattern recognition, color decomposition, and chromatic stability [11].

These processes saw an increase in performance with the advent of deep-learning-based models. In Ref. [12], convolutional neural networks (CNN) automate the process of texture finding, no longer requiring hand-crafted features and achieving better results at motility event classification. Likewise, other WCE domains such as polyp detection [13–15], bleeding [16,17], ulcer detection [18], and celiac disease diagnosis [19], have benefited from the use of CNNs.

Recently, WCE models have thrived with more advanced methods, as the works in Refs. [3,20–24] demonstrate. Attention mechanisms to let the network learn the important features [20], the use of residual connections with the ResNet model [25] along with metric learning with triplet loss (TL) [26] in Ref. [3], and the ability to create deeper and denser models [21] have enabled them to produce more robust and accurate methods. Noteworthy, disease detection in the gastrointestinal tract has greatly gained from recent advances, with CADx systems being explored in bleeding detection, vascular lesions, ulcers, polyp, and tumors [27–29].

Notwithstanding the recent advances, both traditional machine-learning-based methods and the deep-learning variants suffer from the same problems—lack of labeled data and, in some domain-specific tasks like polyp detection, also highly imbalanced classes. This is formalized and analyzed in Refs. [21,30], where the difficulties of producing models that generalize and do not overfit, product of imbalance, low inter-class variance, and high intra-class variance are inspected in detail. Techniques like dropout, L1 or L2 regularization, and sampling mechanisms have also been applied to WCE in an attempt to soften the problems derived from data imbalances and inter-class and intra-class variances, such as overfitting and failing to generalize [31]. Other works, like [3], show that using TL to learn better embeddings also contributes toward obtaining more robust models. Nonetheless, the problem still remains, WCE is tedious to label due to its length, which often means that researchers have low amounts of labeled data to work with. Moreover, several fields must still tackle with huge imbalances within the data.

2.2. Self-supervised learning

Other approaches to tackle low amounts of labeled data, when pseudo-labeled data is available, and class imbalances in downstream tasks are self-supervision methods, of which a wide range of options are available. For instance, a popular architecture choice was autoencoders [32–34], whose dimensionality-reducing capabilities were believed to be useful for SSL. However, it has been demonstrated that they fail to capture rich information [35], focusing only on compressing data. Thus, their capacity to adapt to any future generic task is hindered at best.

In contrast to the former generative method, where the network learns from a single image, contrastive learning trains on multiple examples or instances of the same image to learn the inherent information [36]. One such way to introduce multiple samples of a single image has been by reordering subsections [37]. This type of SSL encourages the network to learn invariant representations, unlike their generative counterparts. Similarly, when the time dimension is available, reordering can be done based on fragments of the input, as done with audio streams [38]. Additional techniques, like rotation, color jittering, blurring, and cropping, can be applied as shown in Refs. [39,40]. The authors propose SimCLR, an architecture based on ResNet [25] that can be trained with multiple contrastive approaches and a new contrastive loss. They provide a simple framework to perform SSL and benchmark the different methods.

More specifically and related to our application, contrastive SSL from videos has been done by predicting the order of a sequence [41–43], object tracking [44–46], and specialized losses [47,48]. In particular, our method resembles the single-view approach of Time-Contrastive Networks [48], which uses metric learning for temporal coherence. Their work, however, diverges from ours because they do not focus on the embeddings' richness nor task-generalization. Given the nature of their action imitation task, they limit their triplets to be in a single sequence and do not explore the embedding quality, whereas our work aims to learn generalized and rich embeddings from hours-long videos, exploring inter-sequence and inter-video triplets, for further usage in downstream tasks.

In medical imagining, some efforts have been made in regards to SSL and semi-supervised training [49,50]. For instance Ref. [51], uses a generative network to create simulated postoperative MRI images, which used in an SSL step obtains better results. Other tasks, such as pneumonia detection and multi-organ segmentation [52], also show improvements by means of SSL-based on samples' patch reordering. Likewise, SSL has also been applied to WCE related tasks, using distortions to the original images in Ref. [53], combined with multi-task learning to detect inflammatory and vascular lesions. Similarly [54], minimizes the difference in predictions between the SSL head and the supervised head, leveraging unlabeled data.

To the best of our knowledge, however, there has been no work that

leverages the temporal aspect of WCE videos, using an SSL process to obtain better representations, which, in turn, would help tackle data-derived problems.

3. Method

An overview of our proposed self-supervised approach is illustrated in Fig. 1. Similar to most methods relying on self-supervised training, our approach is divided into two distinct stages: (a) pretraining a self-supervised network using unlabeled data to obtain rich representations, and (b) finetuning the model using labeled data for a specific task. This section follows the same pattern, explaining both phases first, and finishes by explaining the architecture used.

3.1. Self-supervised pretraining

During the first stage of the process, we aim to extract useful generic information from the unlabeled images, which then can be transferred to deal with many specific tasks by finetuning the model with limited labeled data. In other words, it creates a reduced representation (embedding) of the original image that contains its most important information.

Extracting an embedding can be understood as a process $f(x)$, where a neural network transforms a sample x from the dataset to its compressed and rich representation.

Out of all the possible ways to obtain said embedding, we have chosen to exploit the temporal nature of WCE videos. Our method works by taking sequences of N contiguous frames and creating a relationship between them. Namely, given two frames i, j in the sequence, their relationship is established as the distance $d(i, j)$ between them, counted by the number of frames that separates them.

Unlike the work in Ref. [48], where all samples come from a single sequence, our method must generalize to multiple videos and sequences. Per-frame pseudo-labels are introduced to encode their video identifier along with their position. Given an image i, its pseudo-label is a combination of its video identifier $\gamma(i)$, which can be a simple numbered sequence, and the position inside the video $\delta(i)$, as seen in Equation (1).

$$\bar{y}(i) = M\gamma(i) + \delta(i) \tag{1}$$

Where M must be a large enough number so that $\forall i, M > \delta(i)$. For our particular experiments and datasets, we have chosen $M = 10^6$.

Next, we impose a similarity measure between frames on the sequence so that contrastive learning can be done by finding the inherent relationship between similar and dissimilar images. For that purpose, two images will be consider similar if they are close enough, formalized as $d(i,j) = |\bar{y}(i) - \bar{y}(j)| \leq w$, where $w \leq N$ is a constant chosen beforehand. The pair (i, j) is considered similar (positive) in such cases, and negative otherwise.

In other words, taking a reference image (anchor) in a sequence, all other images within a window of size $2w$ (w images per side) are considered similar. In general, given an N-sequence, all images have between $\min(N, 2w)$ and w positive samples. Images around the edges of the sequences lose up to half the positives, tending towards the latter, while those on the center have the whole spectrum.

The pseudo-labels guarantee that (i, j) negative pairs are consistent with images coming from different videos, as $\gamma(i) \neq \gamma(j)$, thus $d(i, j) \approx |M\gamma(i) - M\gamma(j)| \geq M > w$. Additionally, for two frames i, j extracted from the same video, the formula reduces to the distance in frames between them, $d(i,j) = |\bar{y}(i) - \bar{y}(j)| = |\delta(i) - \delta(j)|$.

Given the above approach to create a similarity measure, the Triplet Loss (TL) [26], a contrastive loss, is introduced to learn the embeddings. TL works by using triplets of samples, where two of the triplet's elements, the anchor a and the positive p, pertain to the same class. The remaining element, the negative n, is of a different class than a. That is, given the embedding of an anchor $f(a)$, a triplet $(f(a), f(p), f(n))$ is formed so that $y(a) = y(p) \neq y(n)$, where $y(\cdot)$ is the class of a sample.

Using Equation (2), TL forces $f(p)$ to be close to $f(a)$ while moving away $f(n)$. It eases the problem by introducing a soft margin α between the positive and negative pairs.

$$TL = max\left(\|f(a) - f(p)\|^2 - \|f(a) - f(n)\|^2 + \alpha, 0\right) \tag{2}$$

Translated to our domain, a triplet is formed by two similar images and a dissimilar image, so that $d(a, p) \leq w$ and $d(a, n) > w$. As shown, TL is directly applicable to WCE videos when used in conjunction with the

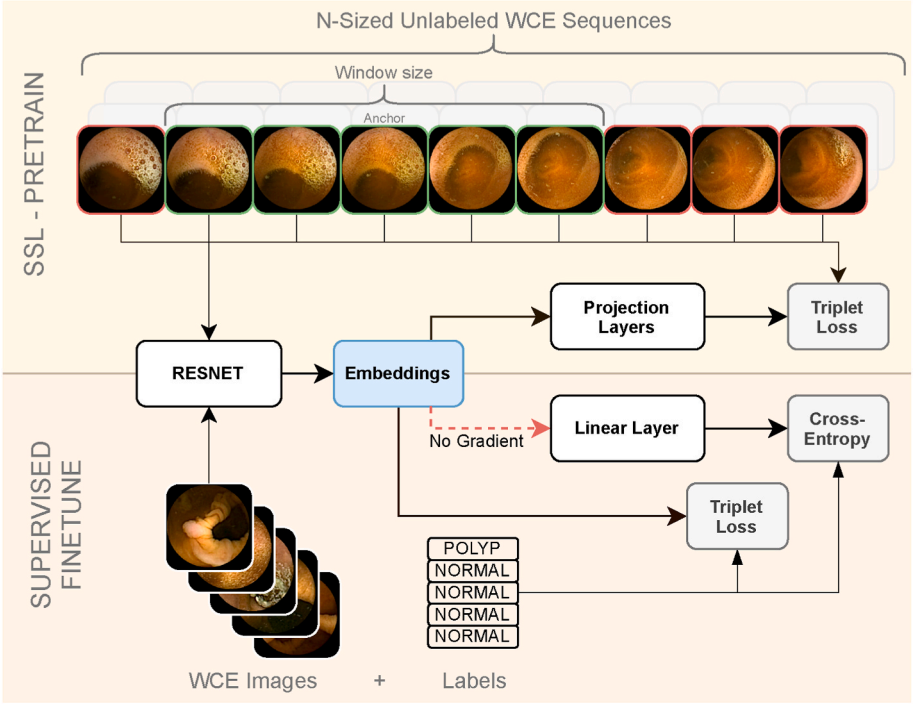

Fig. 1. Overview of the proposed method, including the pretain phase, in the upper half, and the final finetune phase in the lower half.

pseudo-labels, forcing close images in a sequence to have similar representations in the embedding space.

It must be noted that this method is bound to have incorrect pairs, as different videos or sequences could contain similar images, regardless of their distance. Also, WCE videos tend to have periods where the capsule moves at a slow rate, producing many similar images in a relatively long interval, or, on the contrary, moves fast and captures rapidly changing sequences. We estimate those cases to be negligible compared to our dataset's size, being effectively treated as noise during the process.

3.2. Supervised learning

During the second phase of our method, the same model is reused to learn a domain-specific task with limited amounts of data. For instance, the rich representations could be used to model motility events, classify several conditions like bleeding or inflammation, evaluate keyframes, or detect polyps, to name a few.

For that purpose, the process starts with the SSL model's parameters, obtaining embeddings produced by the new dataset and feeding them into a classifier. That classifier needs to access the ground truth labels, as it uses a softmax cross-entropy loss to model the problem.

Following SimCLR findings [39], we have confirmed that fixing the weights obtained during SSL is counterproductive. However, unlike SimCLR, which assumes balanced problems, we use the approach proposed by Laiz et al. [3], where the TL is used to modify the embeddings. As such, the gradient coming from the linear classifier is removed so that it cannot negatively impact the embeddings due to the imbalance. Instead, a TL is imposed on them to facilitate the network to finetune the dataset representations.

However, unlike in the previous step, the TL no longer uses the pseudo-labels created through our method. Triplets are formed by considering the real labels of the images, which are domain-specific and help finetune the embeddings to the particular task. To further reference it and avoid confusion, the term TL_{sup} will be used.

The TL_{sup} is trained in batch all mode, which considers all triplets regardless of their difficulty. No special sampling algorithm is introduced; the only restriction we impose is for a batch to have a proportional representation from all classes. Other than that, data is randomly sampled.

The final loss obtained in this model is the linear combination of both the cross-entropy loss and the triplet loss, as shown in Equation (3).

$$L_{sup} = TL_{sup} + L_{crossentropy} \tag{3}$$

3.3. Architecture

The backbone of our architecture consists of a ResNet-50 [25], as can be seen in Fig. 2a. Most works that extract or require embeddings use the output of the ResNet model directly as their representations, but following the work in SimCLR, we decided to explore the possibility of including several projection layers.

Each projection layer consists of a ReLU activation followed by a dense layer. We restrict all the projection layers to have the same dimensionality, which must be lower than the 2048 given by ResNet. While our pretrain phase benefits from the reduced complexity after the projection, the final finetuning network utilizes the whole 2048-sized embedding to allow for better detection rates. These layers, along with their configuration, hyperparameters, and performance, are studied below. Ultimately, they are found to be beneficial for domain-specific tasks.

Once the pretrain is done, at the beginning of the finetune phase all learned parameters are kept except for the projection layers, which are removed from the model, as can be observed in Fig. 2b. Classification is done through a linear layer (a dense layer without any activation) and a cross-entropy loss. As denoted in red and a dashed line in Fig. 2b, we eliminate the gradient coming from the linear classifier to stop it from modifying the embedding. Only the TL loss is able to tune the representations.

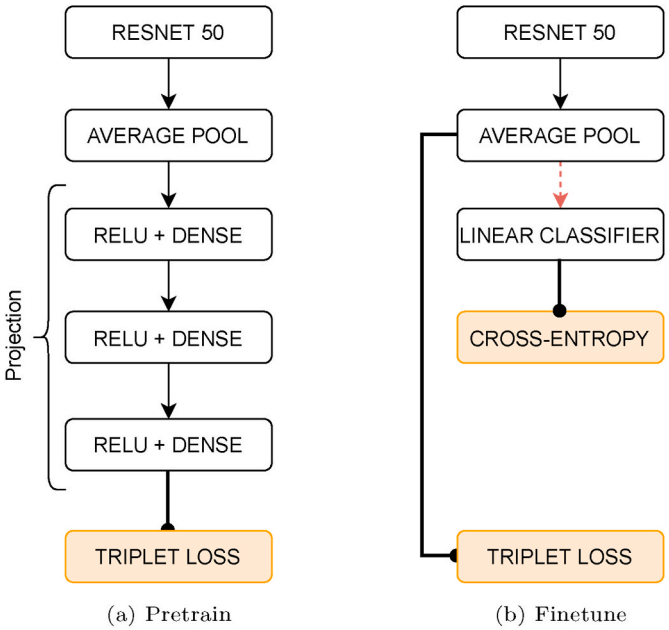

Fig. 2. Detailed network architecture. The parameters obtained during pretain for ResNet are used in the finetune phase, while the projection layers are removed. Here, the dashed red line denotes that gradient is stopped.

It must be remarked that the TL losses used in both phases of the architecture are different. As pointed out, the first phase uses the pseudo-labels deducted from videos, while the second uses the ground truth labels.

4. Discussion and results

This section begins by laying out the datasets used during both steps of the method. Further, it explains the implementation details, such as preprocessing steps and training strategies. A subsection is devoted explicitly to the SSL hyperparameters, justifying and proving the choices made. Finally, individual results are shown for each dataset, discussing the results qualitatively and quantitatively.

4.1. Datasets

Three datasets are used throughout the process. The Generic WCE videos dataset is employed only for the common SSL stage, while the other two are each used to evaluate their own downstream tasks.

4.1.1. Generic WCE videos

This dataset consists of a total of 49 unlabeled WCE videos, each from different patients, obtained with Medtronic PillCam SB2. From those videos, only the small intestine and colon segments are used, selecting a total of 1,185,033 frames.

Even though these images are not labeled, pseudo-labels can be introduced through our method, which makes this dataset suitable for a pretrain step using SSL.

4.1.2. Polyp WCE

The dataset consists of 248,136 frames sampled from 120 procedures performed using Medtronic PillCam SB3 and PillCam Colon 2. Notably, they are not the same videos as the subsection above. Of those frames, 2,080 contain polyps, while 246,056 do not. An initial report is produced by eight expert readers, endoscopy nurses with at least three months of experience, who tag potential polyp frames, and others that

require detailed revision. Then, two medical doctors (one gastroenterologist, and one internal medicine) obtain the final version of the dataset. The polyp's sizes, as reported in Table 2, were obtained through Rapid PillCam Software V9. The largest polyp was determined to be 16 mm. Tumors were considered positive, while any other pathology, like ileal lymphoid hyperplasia, bleeding, and diverticulitis, were discarded from the dataset.

Unlike the Generic WCE videos, this dataset uses SB3 and Colon 2 as sources. It is shown in Ref. [55] that using SB3 from SB2 is possible, while [3] demonstrates that having mixed sources poses no problems for polyp detection.

Overall, this dataset suffers from the exact problems this publication aims to tackle: only 0.85% of all images contain polyps. It is a highly imbalanced problem with an objectively low amount of samples compared to traditional deep learning settings.

4.1.3. CAD-CAP WCE

This public dataset was compiled during the Gastrointestinal Image ANAlysis (GIANA) challenge [56]. It consists of three balanced classes: normal, inflammatory, and vascular lesion, each with approximately 600 images for a total of 1,800 images.

Although the classes are balanced, the total amount of samples is much smaller than the other supervised dataset. Thus, this set can be used to test if the SSL process has captured enough rich information to avoid overfitting.

4.2. Implementation details

We performed all the experiments on one NVIDIA Titan Xp GPU, implementing the entire architecture in TensorFlow 2.4. The backbone network, a ResNet-50, was initialized using the Imagenet trained model, while the projection layers were randomly initialized.

4.2.1. Preprocessing

All data, including the used in pretrain and finetune, was processed using standard data augmentation (DA) techniques during the training phase, such as color jittering, grayscale conversion, and random rotations and flips.

Only RGB channels are used during all stages, keeping the images' size at 256 by 256 pixels and downsizing them using bilinear interpolation without antialias when needed. We also introduced a mask with a radius of 128 pixels to eliminate any artifacts present at the borders of the images, making sure that no specific noise or patterns could identify either a dataset or a particular video.

For our finetune step, as is customary in the field due to the low number of images, the use of DA is mandatory to avoid overfitting. We found that not introducing this same augmentation on the pretrain step negatively affected our final classification results. Thus, all sections below assume the use of DA techniques for training. During evaluation no prepocessing, other than resizing, is done to the data.

Table 1
Hyperparameters tested during the self-supervised training, combining different Sequence Sizes (N) and Window Sizes (w). Resampling indicates that, in a single batch, all sequences come from the same video. Note that resampling only makes sense if N is smaller and multiple of the batch size.

Sequence Size	Sequences per Batch	Window Size	Resample	AUC (%)
9	8	3	No	93.51 ± 1.35
9	8	3	Yes	93.23 ± 1.78
9	8	6	No	93.49 ± 1.31
9	8	6	Yes	93.81 ± 2.12
18	4	3	No	93.68 ± 1.97
18	4	6	No	93.47 ± 1.11
18	4	6	Yes	92.91 ± 2.70
18	4	9	No	93.42 ± 1.62
18	4	9	Yes	93.62 ± 1.63
72	1	6	–	94.12 ± 1.35
72	**1**	**9**	**–**	**94.60 ± 1.15**
72	1	18	–	94.14 ± 2.12
72	1	32	–	94.53 ± 0.96

Table 2
Morphology - Polyp's size in the Polyp WCE dataset, as reported in Ref. [3].

		Morphology			Total
		Sessile	Pedunculated	Undefined	
Size	Small (2–6 mm)	65	4	19	88
	Medium (7–11 mm)	29	4	20	53
	Large (12+ mm)	8	3	13	24
Total		102	11	52	165

4.2.2. Self-supervised learning

The unlabeled Generic WCE videos were used as training data during this stage. The network was optimized using stochastic gradient descent, without momentum, for a total of 21,000 batches with 72 images each (about 2 h and 30 min on our GPU). In our best-performing configuration, the network processes 21,000 sequences. The learning rate was fixed to 0.1, and was divided by 5 every 4,300 iterations. Throughout the process, we used an L2 weight decay of 0.0001. We experimented with multiple values, reaching the same conclusion as SimCLR [39], whereas any low value helps regularize the embedding pre-projection. Finally, we used a batch all strategy for triplet loss, with unnormalized embeddings and a margin of 0.2.

While the SSL network will be used as is, after training with the Generic WCE videos, it is required to find the best set of hyperparameters. To such means, a procedure has been devised. For a particular set of hyperparameters, the network is normally trained, then finetuned over the polyp dataset, and finally evaluated using Area Under the Curve (AUC) computed from Receiver Operating Characteristics (ROC). Here, the polyp dataset is only used as a proxy to evaluate how the hyper-parameters perform and not as a proper evaluation of the downstream task. For instance, this procedure uses a five-fold cross-validation over randomly selected samples from the Polyp WCE videos, whereas the downstream task will be evaluated with complete videos.

4.2.3. Supervised learning

For each of the two supervised datasets, Polyp and CAD-CAP, the entire pretrained network was finetuned with a linear classifier on top of the learned representation. All datasets were equally trained with a learning rate of 0.01, decaying it by 10 every 1,500 iterations for a total of 4,500 steps.

4.3. SSL hyperparameters

We first performed experiments to choose the sequences' length N, window size w, and whether multiple videos should be used in a single batch or not. Due to our available GPU memory, we could fit at most 72 images in a single batch, which set an upper bound to N. We designed several models, see Table 1, to select the best performing combination. Although the results show no statistically significant difference among some, it can be observed that sequences of 72 images, where all images

come from the same video, tend to give better results. Sampling from one video or multiple at once, within a set sequence and window size, has a lower effect on the results than the length of the sequence. Due to hardware limitations, further combinations could not be tested. For instance, it is encouraged to try whether multiple sequences of 72 images are beneficial for a particular downstream task.

Most images will be relatively similar and close when using a continuous stream of 72 images. Therefore, triplets formed for TL will consist of hard negatives, namely from samples that are difficult to distinguish. Oppositely, mixing several short sequences in a single batch will produce negatives that are too easy to distinguish from their anchors.

We believe this added difficulty, albeit making the training process slower, helps the network extract more meaningful information from the images. Thus, richer embeddings are produced, which can then perform better in later downstream tasks. For future experiments, N was fixed to 72, obtained continuously from a single video, and w to 9 images.

Next, we pinpointed the benefits of adding projection layers. We verified, as can be observed in Table 3, whether adding these additional parameters during the pretraining phase yielded better results during polyp detection. It is of particular importance to remark that any projection layer added is then removed during the second phase; thus the same number of parameters is kept regardless of the choices made here.

Particularly, the optimal combination for our particular task seemed to be at 3 layers, each of 128 parameters, which yields a substantial improvement compared to using none and outperforms more complex solutions.

After finding the set of hyper-parameters that performs best, all models used for hyper-parameter evaluation are discarded. Downstream tasks are finetuned with the SSL network trained with the Generic WCE videos dataset, with $N = 72$, $w = 9$, and 3 projection layers with 128 parameters each.

4.4. Results

In this subsection, first the quality of the embeddings learned during the self-supervised learning is evaluated. Then, we explore the results obtained with two downstream specific tasks.

4.4.1. SSL embeddings

As stated, our SSL process aims to learn rich embeddings. To such end, we use the temporal sequences extracted from WCE videos to make the network learn when two images are close or not in the video. It is expected that two embeddings of consecutive images are similar.

Taking into account that euclidean distance is used to measure similarity in the TL function, two embeddings are considered close if their distance is relatively near the margin parameter, or distant otherwise. As can be seen from Fig. 3, the network successfully distinguishes not only images that are completely different but also correctly represents images that are similar while not being consecutive.

Similarly, some samples are close to frames of other videos while maintaining evident similarities, which serves to justify that the network has not learned features specific to a video, but, rather, it has trained for rich information. Our time-based contrastive learning implicitly enables the model to identify similarities between different videos with similar events, which is vital for SSL, as the finetune process needs this augmented information to function properly.

To further validate the embeddings, we obtained a t-SNE representation [57] of one WCE video. As can be seen in Fig. 5a, frames that are visually close, containing similar structures and colors, are densely packed in the same area of the representation. This indicates that their embeddings are also close, verifying that the network has learned our contrastive metric successfully. Likewise, the network has learned that images that are close in the video, are naturally similar, Fig. 5b. The smooth gradient of colors, following the *viridis* scheme, along with the clusters of similar colors, further indicate that similar images have similar embeddings.

Table 3
Study of the effect of adding several projection layers with a varying number of parameters. Each projection layer consists of a ReLU activation followed by a dense layer. All dense layers have the same amount of parameters (dimensionality).

Projection Layers	Projection Dimensionality	AUC (%)
0	–	92.97 ± 1.19
1	128	93.02 ± 1.39
2	128	94.09 ± 1.28
3	**128**	**94.60 ± 1.15**
3	256	93.56 ± 1.53
6	128	93.85 ± 1.80

4.4.2. Polyp dataset

Following previous work from Laiz et al. [3], we abandon traditional metrics used in polyp detection. Accuracy, for instance, is a skewed metric under such data imbalances, favoring the class with most examples in detriment to the overall performance. Thus, as proposed in their publication, we adopt AUC ROC as the primary metric. Moreover, following the same procedure in Ref. [3], sensitivity at set specificity thresholds, namely 95%, 90%, and 80%, are also reported. Not only are they robust towards imbalance, but most importantly, they provide helpful information regarding the number of images a physician needs to check to obtain a certain level of performance in polyp detection. For instance, this metric gives a measure of how many polyps would be detected if a percentage of negatives was discarded based on the classifier.

To ensure that similar images, which are commonly found in sequential frames in videos, are not present in both train and evaluation simultaneously, we split the dataset based on whole videos. Consequently, a patient can only be found either in train or evaluation, but never in both. Failing to do so would overestimate the performance, producing better results while probably failing to generalize with new data.

The baseline for this particular task, further referred to as Imagenet, uses a ResNet-50 preinitialized with Imagenet and trained on this same dataset. Unlike our model, the Imagenet model uses no SSL nor any contrastive loss. A more advanced model, TL_{BA} as trained in Laiz et al. [3], introduces a TL to the previous model. Finally, the state-of-the-art contrastive learning architecture SimCLR [39], is also compared.

Every result, as seen in Table 4, is reported as the mean value and standard deviation obtained from a 5-fold cross-validation. Each evaluation set is done with whole videos, not individual samples. Also, each fold is finetuned and evaluated independently, starting from exactly the same initial values taken from our pretrained network.

Adding any kind of contrastive losses, as can be seen from TL_{BA} and SimCLR in Table 4, already provides a significant boost over the baseline of 9.91% and 10.02% on the AUC score, respectively. Furthermore, our method based on SSL outperforms the former models by 2.24% and 2.06%, respectively, reaching an AUC score of 95.00%. A detailed view of the ROC curve is provided in Fig. 4, where our model can be seen outperforming the rest, achieving higher true positive detections with a lower false positive rate. This significant improvement can be observed across all metrics, meaning SSL and our particular time-based contrastive learning can extract information that remains otherwise hidden or ignored. Of particular interest are the improvements in the sensitivity at different specificity levels, as shown in Table 4. Our method can give a notable increase in the number of polyps correctly classified when discarding varying amounts of negatives.

Another approach to validation, aside from the quantitative analysis above, is to inspect and visualize the results. In other words, performing a qualitative validation of the results by examining where the model is performing correctly and where it is failing. Miss-classified non-polyp images would add more work to the physician due to having to

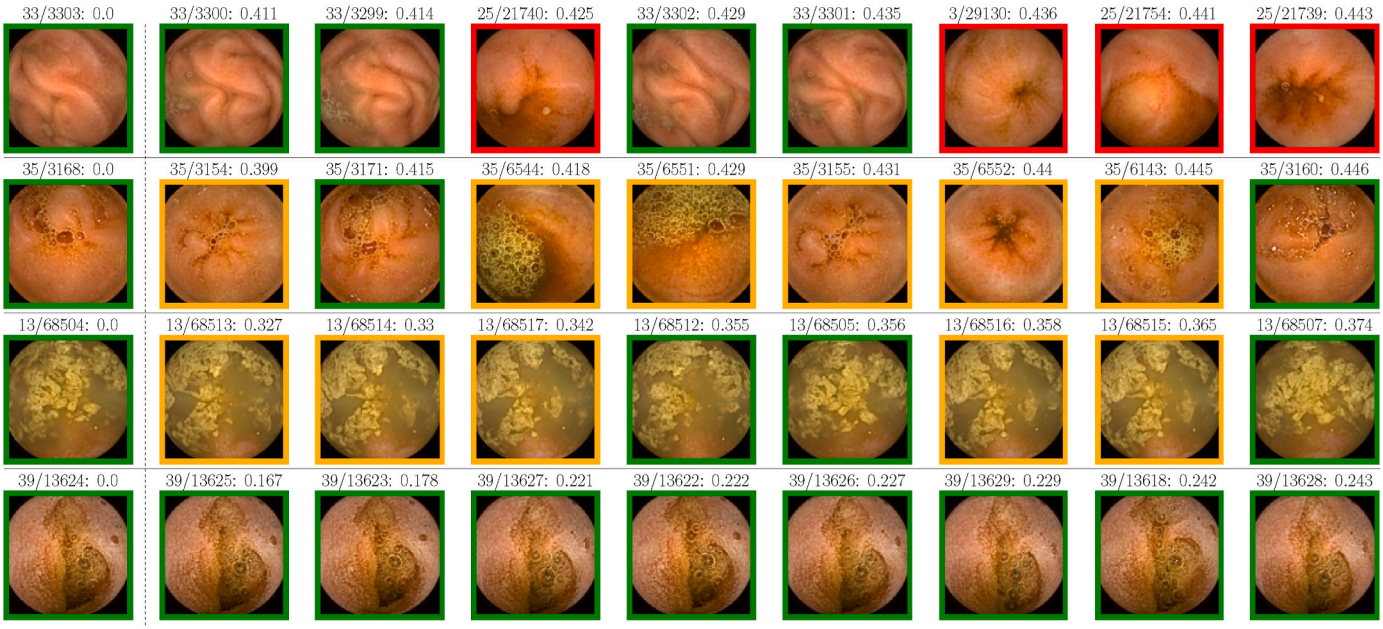

Fig. 3. Given samples from the test set, shown in the first column, each row represents other samples in the set sampled by distance in the embedding space. Each image is titled as *video/frame: distance*, and framed in red if they come from a different video, orange if it is the same video, and green if, additionally to being in the same video, they are within w distance.

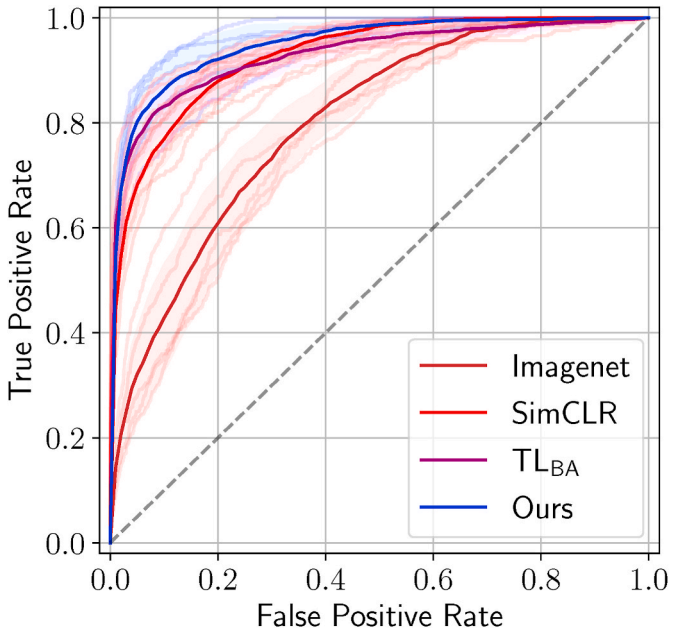

Fig. 4. Receiver Operating Characteristics (ROC) curve for the four models tested for the polyp dataset. Each cross-validation split is shown in lighter versions of its corresponding model color, the mean ROC value is outlined in a darker color, and the standard deviation is provided as the background shade. True Positive Rate indicates the percentage of polyps correctly identified, while False Positive Rate is the percentage of non-polyps misclassified as polyps.

unnecessarily check false positives. However, not showing a polyp frame because the system has falsely classified it as negative can have a devastating effect, with implications more severe than its counterpart case. Fig. 6 depicts two examples of the mentioned cases. It can be seen that the network fails in especially tough cases, where the polyp would be hard to be seen even for a physician. The polyps have been circled for the reader to identify where they are. False positives occur in zones with a more pinkish tone, characteristic of polyps, and always in rugged and wrinkled surfaces, which could explain why the network is mistaking them for polyps.

4.4.3. CAD-CAP dataset

Following the procedure established in Ref. [54], we have split the data into 4 sets and performed a 4-fold cross-validation. As per the original challenge [56], we report in Table 5 the per-class Matthews correlation coefficient (MCC) and F1 scores, and the overall accuracy as p_0.

A naive implementation, using a ResNet-50 and without SSL, fails to correctly classify a significant portion of the data, achieving only a 69.98% accuracy. However, adding SSL to this same model and using the method we propose in this publication immediately boosts every metric by more than 20%. Our implementation reaches a total of 92.77% accuracy without any change to the architecture.

Further, we compare our results with those reported by Guo et al. [54], the current state-of-the-art model for CAD-CAP. They handcrafted a network for this dataset and provide six baselines and one additional model that uses semi-supervision to improve the results. With respect to the baselines, our model obtains higher scores across most metrics, as can be observed in Table 5. We also attain comparable results to their best implementation, which has a semi-supervised phase training over 1807 unlabeled images provided by CAD-CAP that we do not use.

These results, from a clinical point of view, provide a positive step towards the simultaneous detection of several pathologies. For instance, results show that standard models that do not rely on SSL tend to accurately classify normal images, but miss a notable amount of the positive classes. Their SSL counterparts, however, keep the same approximate level of detection for normal samples, while they significantly boost the ability to detect inflammatory and vascular lesions. This encouraging accuracy would enable bringing physicians and experts into the loop, further developing the model and producing a CADx system capable of aiding in diagnosis.

5. Conclusion

In this work, we propose an SSL method that leverages the information in the temporal axis of WCE videos to obtain rich embeddings.

(a) Each embedding is represented with its corresponding image.

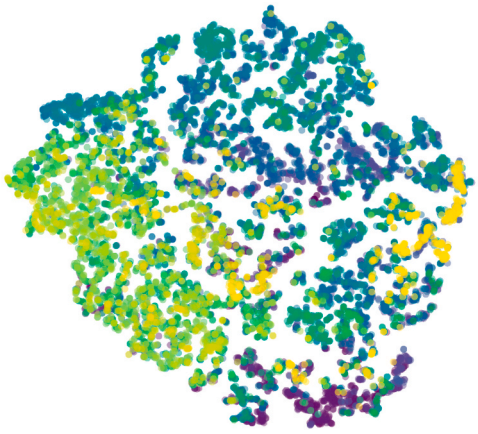

Fig. 5. t-SNE of the embeddings post-projections obtained from one WCE video after the pretrain phase. The representation shows (a) that visually alike images have close embeddings, and (b) that order is preserved.

Table 4
Performance comparison of several methods with the same parameter count. Imagenet refers to a ResNet-50 pretrained on the imagenet dataset and then finetuned with a cross-entropy loss over our dataset. SimCLR has been trained with NT-Xent as per Chen et al. [39]. TL_{BA} is equivalent to Imagenet but trained with an additional triplet loss. Ours is the self-supervised network.

	AUC	Sensitivity %		
Model	**(%)**	**Spec. at 95%**	**Spec. at 90%**	**Spec. at 80%**
Imagenet	82.85 ± 5.72	37.75 ± 9.12	51.49 ± 11.09	66.71 ± 12.15
SimCLR [39]	92.76 ± 1.62	68.13 ± 6.37	76.92 ± 5.40	87.91 ± 3.94
TL_{BA} [3]	92.94 ± 1.87	76.68 ± 4.93	82.86 ± 4.78	88.53 ± 3.76
Ours	**95.00 ± 2.09**	**80.16 ± 6.97**	**86.31 ± 6.20**	**92.09 4.63**

Our method introduces a pseudo-labeling process that enables time-based contrastive learning, forcing frames close in a video to be represented by similar embeddings.

We demonstrate that using this process yields better results in subsequent models specializing in domain-specific tasks. Using the SSL model to classify polyps shows an increase in successful polyp detection, achieving a 95.00% AUC, a significant improvement over existing methods. Similarly, we test the method to detect several events in the GIANA dataset, obtaining comparable results to state-of-the-art models while offering reduced complexity and a more general approach.

It is a limitation of our SSL method that the data used during the pretrain stage must come in a video format. This makes it directly applicable for WCE datasets, but would require adaptation for other medical fields. The pretaining phase is also limited by the hardware capacity, especially so since results show that longer sequences produce richer embeddings. If deployed as a CADx system, our work would only require individual samples and appropriate hardware to run.

Thus, we claim that using SSL when leveraging temporal information is beneficial for WCE models. Most importantly, the method imposes no requirements for the dataset used during the supervised phase, effectively tackling the classical problems commonly encountered in medical imaging: low amounts of data—specially labeled—and severe class imbalances.

Overall, we strongly believe the method is a good step towards better models that empower CADx models in medical interventions. For instance, a higher rate of polyp detection would decrease the time spent by physicians revising WCE videos, allowing for more accurate diagnosis

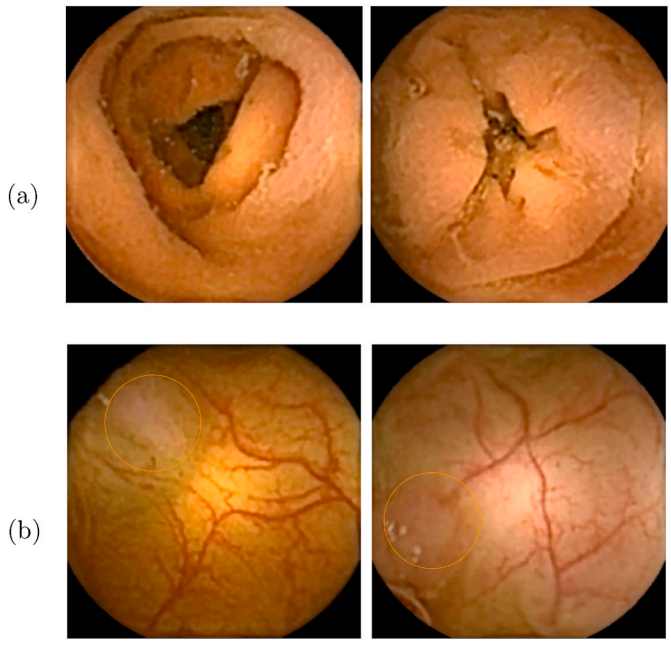

Fig. 6. Random samples from the test set. Row a) shows two false positives, images inaccurately classified as polyps. Row b) depicts two false negatives. The polyps have been circled to help with their identification.

Table 5
Per class and overall results of various methods in GIANA. ResNet is the same architecture as Ours but without the SSL step. Baseline 1 and 6 refer to the baselines reported by Guo et al. [54], while the model with the same name is their semi-supervised performing implementation. Here $\mathbf{p_0}$ indicates the mean accuracy across all classes.

Method	**Class**	**F1-Score (%)**	**MCC (%)**	**p_0 (%)**
ResNet	Normal	73.28 ± 3.57	60.58 ± 5.44	69.98 ± 1.35
	Inflammatory	65.19 ± 2.95	55.86 ± 1.77	
	Vascular	70.79 ± 4.60	65.35 ± 3.80	
Baseline 1 [54]	Normal	94.92 ± 0.71	92.37 ± 1.07	84.99 ± 0.80
	Inflammatory	79.24 ± 1.55	68.72 ± 2.15	
	Vascular	80.75 ± 1.65	71.49 ± 2.57	
Baseline 6 [54]	Normal	96.41 ± 0.84	94.61 ± 1.26	91.92 ± 1.71
	Inflammatory	88.98 ± 2.13	83.44 ± 3.24	
	Vascular	90.27 ± 2.78	85.75 ± 3.73	
Ours	Normal	95.00 ± 1.13	92.57 ± 1.66	92.77 ± 1.20
	Inflammatory	89.87 ± 1.65	84.99 ± 2.46	
	Vascular	90.26 ± 1.76	85.78 ± 2.37	
Guo et al. [54]	Normal	97.41 ± 0.45	96.10 ± 0.69	93.17 ± 1.14
	Inflammatory	90.30 ± 1.56	85.43 ± 2.24	
	Vascular	91.69 ± 1.21	87.78 ± 2.06	

in shorter amounts of time.

Future work could focus on exploring other SSL architectures that might boost the downstream tasks' performance, while exploring other hyper-parameters settings and sampling mechanisms. Moreover, expanding the method to other WCE domains and other medical fields would also be of high interest.

Declaration of competing interest

None declared.

Acknowledgments

This work was partially founded by MINECO Grant RTI2018-095232-B-C21, SGR 1742, Innovate UK project 104633, and by an FPU grant (Formacion de Profesorado Universitario) from the Spanish Ministry of Universities to Guillem Pascual (FPU16/06843). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp Pascal GPU used for this research.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, Nature 405 (6785) (2000) 417–418, https://doi.org/10.1038/35013140.

[2] M. Vasilakakis, A. Koulaouzidis, D.E. Yung, J.N. Plevris, E. Toth, D.K. Iakovidis, Follow-up on: optimizing lesion detection in small bowel capsule endoscopy and beyond: from present problems to future solutions, Expet Rev. Gastroenterol. Hepatol. 13 (2) (2019) 129–141, https://doi.org/10.1080/17474124.2019.1553616.

[3] P. Laiz, J. Vitrià, H. Wenzek, C. Malagelada, F. Azpiroz, S. Seguí, WCE polyp detection with triplet based embeddings, Comput. Med. Imag. Graph. 86 (October) (2020), https://doi.org/10.1016/j.compmedimag.2020.101794 arXiv: 1912.04643.

[4] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: generative or contrastive, IEEE Trans. Knowl. Data Eng. (2021), https://doi.org/10.1109/TKDE.2021.3090866, 01, 1–1. arXiv:2006.08218.

[5] Y. Fu, W. Zhang, M. Mandal, M.Q. Meng, Computer-aided bleeding detection in WCE video, IEEE J Biomed.Health Inf. 18 (2) (2014) 636–642, https://doi.org/10.1109/JBHI.2013.2257819.

[6] G. Lv, G. Yan, Z. Wang, Bleeding detection in wireless capsule endoscopy images based on color invariants and spatial pyramids using support vector machines, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, vol. 2011, Annu Int Conf IEEE Eng Med Biol Soc, 2011, pp. 6643–6646, https://doi.org/10.1109/IEMBS.2011.6091638.

[7] Y. Yuan, M.Q. Meng, Automatic bleeding frame detection in the wireless capsule endoscopy images, in: Proceedings - IEEE International Conference on Robotics and Automation, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 1310–1315, https://doi.org/10.1109/ICRA.2015.7139360, 2015-June.

[8] K. Pogorelov, S. Suman, F. Azmadi Hussin, A. Saeed Malik, O. Ostroukhova, M. Riegler, P. Halvorsen, S. Hooi Ho, K.L. Goh, Bleeding detection in wireless capsule endoscopy videos — color versus texture features, J. Appl. Clin. Med. Phys. 20 (8) (2019) 141–154, https://doi.org/10.1002/acm2.12662.

[9] L.A. Alexandre, J. Casteleiro, N. Nobre, Polyp detection in endoscopic video using SVMs, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 4702 LNAI, Springer Verlag, 2007, pp. 358–365, https://doi.org/10.1007/978-3-540-74976-9_34.

[10] J.-Y. Yeh, T.-H. Wu, W.-J. Tsai, Bleeding and ulcer detection using wireless capsule endoscopy images, J. Software Eng. Appl. 7 (5) (2014) 422–432, https://doi.org/10.4236/jsea.2014.75039.

[11] C. Malagelada, F. De Iorio, F. Azpiroz, A. Accarino, S. Segui, P. Radeva, J. R. Malagelada, New insight into intestinal motor function via noninvasive endoluminal image analysis, Gastroenterology 135 (4) (2008) 1155–1162, https://doi.org/10.1053/j.gastro.2008.06.084.

[12] S. Seguí, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, J. Vitrià, Generic feature learning for wireless capsule endoscopy analysis, Comput. Biol. Med. 79 (October) (2016) 163–172, https://doi.org/10.1016/j.compbiomed.2016.10.011.

[13] D.K. Iakovidis, S.V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, V. P. Plagianakos, Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification, IEEE Trans. Med. Imag. 37 (10) (2018) 2196–2210, https://doi.org/10.1109/TMI.2018.2837002.

[14] T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara, T. Matsuda, S. Tanaka, K. Koike, T. Tada, Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network, Gastrointest. Endosc. 89 (2) (2019) 357–363, https://doi.org/10.1016/j.gie.2018.10.027, e2.

[15] E.S. Nadimi, M.M. Buijs, J. Herp, R. Kroijer, M. Kobaek-Larsen, E. Nielsen, C. D. Pedersen, V. Blanes-Vidal, G. Baatrup, Application of deep learning for autonomous detection and localization of colorectal polyps in wireless colon capsule endoscopy, Comput. Electr. Eng. 81 (2020), 106531, https://doi.org/10.1016/j.compeleceng.2019.106531.

[16] A. Caroppo, A. Leone, P. Siciliano, Deep transfer learning approaches for bleeding detection in endoscopy images, Comput. Med. Imag. Graph. 88 (April 2020) (2021), 101852, https://doi.org/10.1016/j.compmedimag.2020.101852.

[17] M.A. Khan, S. Kadry, M. Alhaisoni, Y. Nam, Y. Zhang, V. Rajinikanth, M.S. Sarfraz, Computer-aided gastrointestinal diseases analysis from wireless capsule endoscopy: a framework of best features selection, IEEE Access 8 (2020) 132850–132859, https://doi.org/10.1109/ACCESS.2020.3010448.

[18] V. V, K.V. Prashanth, Ulcer detection in Wireless Capsule Endoscopy images using deep CNN, J. King Saud Univ.. Comput. Inf.. Sci. (sep 2020), https://doi.org/10.1016/j.jksuci.2020.09.008.

[19] X. Wang, H. Qian, E.J. Ciaccio, S.K. Lewis, G. Bhagat, P.H. Green, S. Xu, L. Huang, R. Gao, Y. Liu, Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction, Comput. Methods Progr. Biomed. 187 (2020), 105236, https://doi.org/10.1016/j.cmpb.2019.105236.

[20] S. Jain, A. Seal, A. Ojha, A. Yazidi, J. Bures, I. Tacheci, O. Krejcar, A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images, Comput. Biol. Med. 137 (2021), 104789, https://doi.org/10.1016/j.compbiomed.2021.104789.

[21] Y. Yuan, W. Qin, B. Ibragimov, G. Zhang, B. Han, M.Q. Meng, L. Xing, Densely connected neural network with unbalanced discriminant and category sensitive constraints for polyp recognition, IEEE Trans. Autom. Sci. Eng. 17 (2) (2020) 574–583, https://doi.org/10.1109/TASE.2019.2936645.

[22] A.K. Kundu, S.A. Fattah, Probability density function based modeling of spatial feature variation in capsule endoscopy data for automatic bleeding detection, Comput. Biol. Med. 115 (2019), 103478, https://doi.org/10.1016/j.compbiomed.2019.103478.

[23] S. Jain, A. Seal, A. Ojha, O. Krejcar, J. Bureš, I. Tachecí, A. Yazidi, Detection of abnormality in wireless capsule endoscopy images using fractal features, Comput. Biol. Med. 127 (2020), 104094, https://doi.org/10.1016/j.compbiomed.2020.104094.

[24] X. Guo, Z. Chen, J. Liu, Y. Yuan, Non-equivalent images and pixels: confidence-aware resampling with meta-learning mixup for polyp segmentation, Med. Image Anal. 78 (2022) 102394, https://doi.org/10.1016/j.media.2022.102394.

[25] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90, 2016-Decem, arXiv:1512.03385.

[26] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, Adv. Neural Inf. Process. Syst. 16 (2004) 41–48.

[27] R. Trasolini, M.F. Byrne, Artificial Intelligence and Deep Learning for Small Bowel Capsule Endoscopy, jan 2021, https://doi.org/10.1111/den.13896.

[28] O. Attallah, M. Sharkas, GASTRO-CADx: a three stages framework for diagnosing gastrointestinal diseases, PeerJ Computer Science 7 (2021) 1–36, https://doi.org/10.7717/peerj-cs.423.

[29] P. Gilabert, A. Watson, H. Wenzek, Artificial Intelligence to Improve Polyp Detection and Screening Time in Colon Capsule Endoscopy, Scientific Reports, In Review (jan 2022). doi:10.21203/RS.3.RS-1278962/V1.

[30] A. Akay, H. Hess, Deep learning: current and emerging applications in medicine and technology, IEEE J Biomed.Health Inf. 23 (3) (2019) 906–920, https://doi.org/10.1109/JBHI.2019.2894713.

[31] S.H. Kim, Y.J. Lim, Artificial Intelligence in Capsule Endoscopy: A Practical Guide to its Past and Future Challenges, sep 2021, https://doi.org/10.3390/diagnostics11091722.

[32] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence, 2013, pp. 399–421, https://doi.org/10.1016/B978-1-4832-1446-7.50035-2.

[33] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, 2014 arXiv:1312.6114.

[34] G.E. Hinton, A. Krizhevsky, S.D. Wang, Transforming auto-encoders, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6791 LNCS, Springer, Berlin, Heidelberg, 2011, pp. 44–51, https://doi.org/10.1007/978-3-642-21735-7_6.

[35] Y. Bengio, Learning deep architectures for AI, Foundations and Trends in Machine Learning 2 (1) (2009) 1–27, https://doi.org/10.1561/2200000006.

[36] W. Falcon, K. Cho, A Framework for Contrastive Self-Supervised Learning and Designing A New Approach, arXiv (aug 2020). arXiv:2009.00104.

[37] I. Misra, L. van der Maaten, Self-supervised learning of pretext-invariant representations, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2020, pp. 6706–6716, https://doi.org/10.1109/CVPR42600.2020.00674, arXiv: 1912.01991.

[38] A. van den Oord, Y. Li, O. Vinyals, Representation Learning with Contrastive Predictive Coding, jul 2018 arXiv:1807.03748.

[39] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A Simple Framework for Contrastive Learning of Visual Representations, arXiv (Figure 1) (2020). arXiv:2002.05709.

[40] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big Self-Supervised Models Are Strong Semi-supervised Learners, 2020, pp. 1–18, arXiv (NeurIPS), arXiv:2006.10029.

[41] I. Misra, C. Lawrence Zitnick, M. Hebert, Shuffle and learn: unsupervised learning using temporal order verification, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in

Bioinformatics), 9905 LNCS, Springer, Cham, 2016, pp. 527–544, https://doi.org/10.1007/978-3-319-46448-0_32, arXiv:1603.08561.
[42] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, Y. Zhuang, Self-supervised spatiotemporal learning via video clip order prediction, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 10326–10335, https://doi.org/10.1109/CVPR.2019.01058, 2019-June.
[43] H.Y. Lee, J.B. Huang, M. Singh, M.H. Yang, Unsupervised representation learning by sorting sequences, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 667–676, https://doi.org/10.1109/ICCV.2017.79, 2017-Octob, arXiv:1708.01246.
[44] D. Pathak, R. Girshick, P. Dollár, T. Darrell, B. Hariharan, Learning features by watching objects move, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 6024–6033, https://doi.org/10.1109/CVPR.2017.638, 2017-Janua, arXiv:1612.06370.
[45] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2794–2802, https://doi.org/10.1109/ICCV.2015.320, 2015 Inter, arXiv:1505.00687.
[46] X. Wang, A. Jabri, A.A. Efros, Learning correspondence from the cycle-consistency of time, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2019, pp. 2561–2571, https://doi.org/10.1109/CVPR.2019.00267, 2019-June, arXiv:1903.07593.
[47] M. Tschannen, J. Djolonga, M. Ritter, A. Mahendran, N. Houlsby, S. Gelly, M. Lucic, Self-supervised learning of video-induced visual invariances, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2794–2802, arXiv:1910.04867.
[48] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, G. Brain, Time-contrastive networks: self-supervised learning from video, in: Proceedings - IEEE International Conference on Robotics and Automation, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1134–1141, https://doi.org/10.1109/ICRA.2018.8462891, arXiv:1704.06888.
[49] V. Cheplygina, M. de Bruijne, J.P. Pluim, Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, Med. Image Anal. 54 (2019) 280–296, https://doi.org/10.1016/j.media.2019.03.009, arXiv:1804.06353.
[50] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, M. Norouzi, Big Self-Supervised Models Advance Medical Image Classification, vol. 1, jan 2021 arXiv:2101.05224.
[51] F. Pérez-García, R. Dorent, M. Rizzi, F. Cardinale, V. Frazzini, V. Navarro, C. Essert, I. Ollivier, T. Vercauteren, R. Sparks, J.S. Duncan, S. Ourselin, A self-supervised learning strategy for postoperative brain cavity segmentation simulating resections, Int. J. Comput. Assist. Radiol. Surg. 82 (2021) 1–9, https://doi.org/10.1007/s11548-021-02420-2, arXiv:2105.11239.
[52] F. Navarro, C. Watanabe, S. Shit, A. Sekuboyina, J. C. Peeken, S. E. Combs, B. H. Menze, Evaluating the Robustness of Self-Supervised Learning in Medical Imaging (May 2021). arXiv:2105.06986.
[53] A. Vats, M. Pedersen, A. Mohammed, Ø. Hovde, Learning More for Free - A Multi Task Learning Approach for Improved Pathology Classification in Capsule Endoscopy, arXiv (Jun 2021). arXiv:2106.16162, doi:10.1007/978-3-030-87234-2_1.
[54] X. Guo, Y. Yuan, Semi-supervised WCE image classification with adaptive aggregated attention, Med. Image Anal. 64 (2020) 101733, https://doi.org/10.1016/j.media.2020.101733.
[55] P. Laiz, J. Vitria, S. Segui, Using the triplet loss for domain adaptation in WCE, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, 2019, pp. 399–405, https://doi.org/10.1109/ICCVW.2019.00051.
[56] X. Dray, C. Li, J. Saurin, F. Cholet, G. Rahmi, J. Le Mouel, C. Leandri, S. Lecleire, X. Amiot, J. Delvaux, C. Duburque, R. Gérard, R. Leenhardt, F. Mesli, G. Vanbiervliet, I. Nion Larmurier, S. Sacher-Huvelin, C. Simon-Chane, R. Olivier, A. Histace, CAD-CAP: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle, in: Journées Francophones d'Hépato-Gastroentérologie et d'Oncologie Digestive (JFHOD), vol. 50, Georg Thieme Verlag KG, 2018, https://doi.org/10.1055/s-0038-1623358, 000441.
[57] L. Van Der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2625.