





Article

Ensembles of Convolutional Neural Networks and Transformers for Polyp Segmentation

Loris Nanni ^{1,*} , Carlo Fantozzi ¹ , Andrea Loreggia ²  and Alessandra Lumini ³ ¹ Department of Information Engineering, University of Padova, 35122 Padova, Italy; carlo.fantozzi@unipd.it² Department of Information Engineering, University of Brescia, 25121 Brescia, Italy³ Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy

* Correspondence: loris.nanni@unipd.it

Abstract: In the realm of computer vision, semantic segmentation is the task of recognizing objects in images at the pixel level. This is done by performing a classification of each pixel. The task is complex and requires sophisticated skills and knowledge about the context to identify objects' boundaries. The importance of semantic segmentation in many domains is undisputed. In medical diagnostics, it simplifies the early detection of pathologies, thus mitigating the possible consequences. In this work, we provide a review of the literature on deep ensemble learning models for polyp segmentation and develop new ensembles based on convolutional neural networks and transformers. The development of an effective ensemble entails ensuring diversity between its components. To this end, we combined different models (HarDNet-MSEG, Polyp-PVT, and HSNet) trained with different data augmentation techniques, optimization methods, and learning rates, which we experimentally demonstrate to be useful to form a better ensemble. Most importantly, we introduce a new method to obtain the segmentation mask by averaging intermediate masks after the sigmoid layer. In our extensive experimental evaluation, the average performance of the proposed ensembles over five prominent datasets beat any other solution that we know of. Furthermore, the ensembles also performed better than the state-of-the-art on two of the five datasets, when individually considered, without having been specifically trained for them.

Keywords: polyp segmentation; computer vision; ensemble; transformers; convolutional neural networks



Citation: Nanni, L.; Fantozzi, C.; Loreggia, A.; Lumini, A. Ensembles of Convolutional Neural Networks and Transformers for Polyp Segmentation. *Sensors* **2023**, *23*, 4688. <https://doi.org/10.3390/s23104688>

Academic Editor: Alexander Wong

Received: 13 March 2023

Revised: 29 April 2023

Accepted: 9 May 2023

Published: 12 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Colon polyps are among the preliminary manifestations of colorectal cancer, one of the cancers with the highest incidence [1]. The identification of precancerous polyps is essential during screening, as early detection and accurate diagnosis are the keys to effective treatment and low mortality [2]. Each 1% increase in polyp detection reduces the incidence of colon cancer by approximately 3% [3]. Currently, colonoscopy is the gold standard adopted in clinical practice to detect diseased tissue in the gastrointestinal tract. However, the accuracy of the actual match depends on the physician's skill and requires a great deal of effort. Therefore, for the clinical prevention of colorectal cancer, it is crucial to have automatic methods that can point out all existing polyps with high accuracy. Artificial intelligence and machine learning models have been widely applied to the semantic segmentation of polyps in medical images. Two examples of colonoscopy images of polyps and their segmentations are shown in Figure 1.

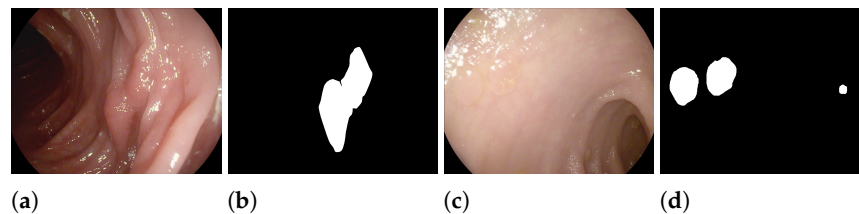


Figure 1. Two examples of the content of the ETIS-Larib [4] dataset for the semantic segmentation of polyps: (a,c) original images; (b,d) ground truth.

Traditional approaches to segmentation have used, for example, geometric analysis and a frame-based model [5] or a hybrid context–shape approach [6]. Such approaches are hardly capable of extracting global context information and are not robust in complex scenes, mainly because they rely on hand-crafted features [7]. Deep learning has brought remarkable progress in the field of semantic segmentation: we refer the reader to [8] for a modern survey on the topic. Recently, deep networks have been applied to automatic polyp segmentation in colonoscopy images [9–11].

Regardless of the approach, the well-known “no free lunch” theorem for machine learning highlights that there cannot be a single model that works well on all datasets. Based on this evidence, an effective procedure is to adopt sets (ensembles) of classifiers, often shallow or weak, whose predictions are aggregated to form the output of the system. In an ensemble, individual classifiers are trained so that each generalizes differently in the training space. Ensembles provide state-of-the-art results in many domains, but it is important to secure some properties. One of them is to enforce some kind of diversity in the set of classifiers.

In this scenario, with this work, we provide two main contributions.

- An up-to-date review of ensembles for polyp segmentation. Actually, to the best of our knowledge, no earlier review on the topic is available in the open literature.
- A new ensemble for semantic segmentation based on the HarDNet-MSEG [12], Polyp-PVT [13], and HSNNet [14] network topologies. The empirical evaluation showed that the performance for polyp segmentation was better than the state-of-the-art.

The idea behind the ensemble is to smooth the contribution of the fine-tuning of the hyperparameters for a specific dataset while averaging the performance of a model to deal with multiple domains [15]. We remark that the common practice of using an activation function, such as sigmoid, as the final layer of the model, followed by normalization, actually goes against this purpose because it can cause the average or sum rule to become too similar to a voting rule, thereby reducing the benefits of ensemble design. On the contrary, our ensemble is scientifically relevant because it introduces an approach to obtain smoother intermediate masks that are more suitable to be aggregated for the final segmentation. In turn, this approach shows that a better way to add transformers (part of Polyp-PVT and HSNNet, tested in this work) to an ensemble is to modify the way the final segmentation mask is obtained. The ensemble also provides evidence that applying different approaches to the learning rate strategy is a viable method to build a set of segmentation networks. Furthermore, we demonstrate that a fusion of different convolutional and transformer topologies can achieve state-of-the-art performance.

The paper is organized as follows. Section 2 contributes a review of the literature on ensembles for polyp segmentation. Full details about our ensemble are provided in Section 3. Section 4 illustrates the results of our experiments and provides a comparison with the state-of-the-art. The paper concludes with Section 5, which contains some final remarks and outlines some research opportunities for the future.

2. Related Work

Several researchers have faced the problems of automatic polyp detection and segmentation. As in other domains, early attempts (e.g., [7]) focused on detection and were

based on features extracted manually. More recently, deep networks have been applied to tasks, and interest has shifted toward segmentation [9–11,16]. Deep networks, together with other techniques [17] (Section B), are also in use for image enhancement [18]. In recent years, encoder–decoder architectures have gained popularity, with the most-widely adopted network in this family being U-Net [19]. Other popular networks adopted for polyp segmentation are, reportedly [20], DeepLabv3+, SegNet, FCNs, DeconvNet, PSPNet, and Mask R-CNN. All of them have been proposed in other domains. Several public datasets are available to researchers to train their networks and compare the performance. A recent summary of such datasets can be found in [21]. Details about the datasets we used in our experiments are provided in Section 3.3. In addition to segmentation accuracy, commonly measured by the Dice coefficient and the intersection over union (IoU) (definitions provided in Section 3.2), a performance metric considered by some works is segmentation time because the ultimate goal is to segment all frames of a colonoscopy video to identify as many precancerous polyps as possible and complete the job as fast as possible. All authors customarily report the mean value of the metrics over all the images in the test set, which is the same approach we followed in this paper.

It is essential to mention that state-of-the-art results are obtained by resorting to methods that increase performance beyond the level attained by baseline networks. Two methods are almost universally adopted in the realm of polyp detection and segmentation, as well as in other medical and nonmedical domains: data augmentation and ensemble techniques:

- Data augmentation [22] increases the size of the training set by adding synthetic samples. Such samples can be created in many ways: the most-common approach in computer vision is to generate new images by simply altering existing ones, for instance by flipping, cropping, or rotating them. However, other approaches are possible, including the generation of completely artificial images [23].
- Ensemble techniques [24] increase accuracy by combining the responses of different classifiers (per-pixel classifiers, in the case of semantic segmentation). As in the case of data augmentation, many different solutions have been proposed to combine the answers and to build the classifiers themselves.

Given the importance of ensemble techniques for this work, we now report on the literature on ensembles for polyp segmentation. To the best of our knowledge, 15 published works on the topic have been published. The salient features of these works are summarized in Table 1 (structure of the proposed ensembles), Table 2 (datasets used), Table 3 (performance metrics adopted), and Table 4 (reported performance). Among the 15 papers, we can report the following:

- Two [25,26] do not provide performance figures on public datasets. The source code is not available as well; hence, any comparison is beyond the bounds of possibility.
- Four [27–30] have been superseded by newer publications from the same authors, which show better performance on a wider range of datasets.
- Three [31–33] exhibit results that no longer represent the state-of-the-art for popular datasets, the more so considering that such results were obtained with more benevolent experimental protocols (e.g., ensembles trained and tested on the same dataset) than the one [10] currently adopted by several researchers, including ourselves.
- Four [20,34–36] report outstanding performance figures, but again, they were obtained with less stringent and/or incompletely documented protocols. Except for [36], the source code is not available.

The remaining two works [37,38] were included in the comparisons made during our experiments, as detailed in Section 4. We now briefly summarize, in chronological order, what we consider to be the most relevant of the 15 works on ensembles for polyp segmentation.

In [20], the polyp segmentation method was based on an ensemble built from three different networks: U-Net, SegNet, and PSPNet. The final segmentation was obtained

through per-pixel weighted voting of the outputs of the three networks. The weights were proportional to the performance of the networks measured in the validation phase. Training, validation, and testing were performed with images from three public datasets: CVC-ColonDB [7] (300 images), CVC-ClinicDB [39] (612 images), and ETIS-Larib [4] (196 images). The training phase relied on transfer learning and data augmentation (scaling, flipping, rotations at different angles, and changes in brightness).

In [31], an ensemble of two Mask R-CNN models with different encoding backbones (ResNet-50 and ResNet-101 [40]) was proposed. The final segmentation was the bitwise combination, using the union operator, of the outputs of the two subnetworks. Transfer learning was adopted: the networks were pre-trained on the COCO dataset and fine-tuned with images from the same three datasets considered in [20]. As in [20], data augmentation (scaling, flipping, cropping, padding, random rotations, random shearing, random Gaussian blurring, random contrast normalization, and random changes in brightness) was adopted during training. The authors stated that this was an effective tool for improving segmentation performance, confirming the common conclusion in the literature.

In [34], the authors addressed the segmentation of multiple anatomical structures, including polyps. The proposed segmentation ensemble combines three DeepLabv3+ [41] variants trained with images at different resolutions and with different dilation strides. The authors stated that the ensemble thus captures information at multiple scales. Furthermore, a novel loss function was adopted that is a combination of the cross-entropy loss and the Dice loss. The discussion seems to imply that the outputs of the three networks were combined with a simple max function, that is if any of the three networks says that a given pixel belongs to a polyp, then this is the final output. The ensemble was trained and tested with images from the CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib datasets. The datasets were augmented with standard geometric alterations (reflection, random cropping, translation rotation), elastic distortions, contrast normalization, and boundary enhancement.

In [32], the segmentation network combines the predictions of two U-Net models with ResNet-34 [40] and EfficientNet B2 [42] as their backbones; the details of how the outputs of the two models are combined to provide the final segmentation were not reported. The networks were trained with transfer learning, with initial weights obtained from ImageNet and fine-tuning performed with the publicly available Kvasir-SEG dataset [43] (1000 images). The segmentation accuracy was then tested with 160 images from the MediEval 2020 challenge. As in the previous works, the training phase leveraged data augmentation (scaling, flipping, random rotations, affine transformations, elastic deformations, CutMix regularization, random changes in contrast, and addition of Gaussian noise). The authors observed that CutMix regularization alone, which substitutes a block of pixels in a training image with a random patch from another image in the training batch, increased the accuracy by up to 3% in the validation set.

In [35], different networks (namely, MobileNet, ResNet, and EfficientNet [42]) were first tested as the backbones of U-Net, finding that EfficientNet provided the highest performance. Then, a new ensemble was proposed that combines the segmentation results of two U-Nets with EfficientNet B4 and EfficientNet B5 as the backbones. The outputs of the networks are combined asymmetrically: the output of the second (i.e., the one based on EfficientNet B5) is taken into account for a pixel only if its confidence level that such a pixel belongs to a polyp is greater than 0.96. A novel loss function was adopted during training to take into account the fact that the data were unbalanced, that is the number of non-polyp pixels is much higher than that of polyp pixels. The proposed function is a combination of the standard cross-entropy and asymmetric F_{β} loss functions. As in all the papers mentioned, data augmentation (random scaling, cropping, padding, flipping, random rotations, random shearing, Gaussian blurring, random contrast normalization, and random changes in brightness) was employed during training. Transfer learning was also applied: the initial U-Net model was pre-trained on the ImageNet dataset and

fine-tuned on the CVC-ClinicDB dataset. The performance of the ensemble was evaluated on the CVC-ColonDB and ETIS-Larib datasets.

Finally, we briefly mention [44], since the authors used the term “ensemble” to describe the networks they examined. However, what they did was test the variations of the U-Net architecture with different encoders. Each tested model was a single network, not an ensemble, with a different feature extractor. Based on the results of the experiments, the best-performing feature extractors were DenseNet169 and InceptionResNetV2. As in several other studies, extensive data augmentation (flipping, blurring, sharpening, random changes in contrast and brightness) was applied during training.

All the aforementioned works combined the outputs of two or three networks. A recent line of research has been the exploration of the accuracy advantages of bigger ensembles, whose predictions may be combined hierarchically or in some other complex fashion. In [36], two different ensembles for the semantic segmentation of polyps were discussed. The first ensemble, named TriUNet by the authors, combines three U-Net networks. The second ensemble, called DivergentNets, combines TriUNet with UNet++ [45], FPN [46], DeepLabv3, and DeepLabv3+. The final segmentation mask was an average of the five masks provided by these networks. DivergentNets can be considered an ensemble of size eight, albeit the outputs of the three networks were pre-combined in TriUNet. In [29], several ensembles were tested whose components differed in the backbones adopted, the loss functions, and the optimizers used in the training phase. The base architectures for the networks that made up the ensembles were DeepLabv3+ and HarDNet-MSEG [12]. The size of the ensembles ranged from 2 to 60. The ensembles were trained on 1450 images taken from the Kvasir-SEG and CVC-ClinicDB datasets, then tested on the remaining images from the same datasets (100 from Kvasir-SEG, 62 from CVC-ClinicDB), as well as on three “unseen” datasets: CVC-ColonDB, ETIS-Larib, and the test set from CVC-EndoSceneStill [47]. This is an experimental protocol that was first introduced in [10]. In [37], the authors proposed ensembles of the DeepLabv3+, HarDNet-MSEG, and Polyp-PVT [13] networks, trained with different loss functions and data augmentation methods. A wider range of loss functions was considered than in [29], including weighted combinations of base functions, and more than ten data augmentation techniques were applied in the training phase. The size of the ensembles ranged from 2 to 14. The datasets used for training and testing were the same as in [10]. In [26], an ensemble was proposed that combines the predictions of EffUNet [48], nnU-Net [49] and a hierarchical multiscale attention network [50]. The training set was the one provided by the EndoCV2022 polyp segmentation sub-challenge, with the addition of images from CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib. The training set was manually curated by the authors to remove images with implausible annotations; it is not publicly available. Moreover, the only performance data provided by the authors were on “folds” of data that do not have a documented relationship with public datasets. In [38], the ensemble was made up of two different sub-ensembles, once again based on DeepLabv3+ (backbone: ResNet-101) and HarDNet-MSEG (backbone: HarDNet-68), respectively. Inside each sub-ensemble, diversity was provided by varying the activation functions (15 different loss functions were considered) and the data augmentation strategies. Additionally, in the sub-ensemble based on DeepLabv3+, polyps identified by different networks were not allowed to overlap. The training and testing protocols were, once again, those introduced in [10].

Table 2. Ensembles for polyp segmentation: datasets used. Some names are abbreviated; see Section 3.3. Datasets can be used for training, testing, or both. Datasets from other domains used as a consequence of transfer learning are not listed.

Paper	ColDB	ClinDB	CVC-T	EDD 2020	Polyp-Gen	ETIS	Hyper-Kvasir	Kvasir	MediEval 2020
Guo2019 [20]	✓	✓				✓			
Kang2019 [31]	✓	✓				✓			
Nguyen2019 [34]	✓	✓				✓			
Shrestha2020 [32]								✓	✓
ThuHong2020 [35]	✓	✓				✓			
Hong2021 [33]					✓				
Lumini2021a [27]								✓	
Lumini2021b [28]								✓	
Nanni2021 [29]	✓	✓	✓			✓		✓	
Thambawita2021 [36]					✓		✓		
Tomar2021 [25]					✓				
Nanni2022a [37]	✓	✓	✓			✓		✓	
Nanni2022b [30]	✓	✓	✓			✓		✓	
Tran2022 [26]	✓	✓		✓	✓	✓			
Nanni2023 [38]	✓	✓	✓			✓		✓	

Table 3. Ensembles for polyp segmentation: performance metrics. “FPS” stands for “frames per second”. The authors of [25] provided only a score specific to the EndoCV 2021 Segmentation Generalization Challenge, named “generalization score”.

Paper	Accuracy	Dice	F2	FPS	IoU	Precision	Recall	Sensitivity	Specificity
Guo2019 [20]		✓			✓				
Kang2019 [31]					✓		✓		
Nguyen2019 [34]	✓	✓				✓	✓	✓	✓
Shrestha2020 [32]	✓	✓	✓	✓	✓	✓	✓		
ThuHong2020 [35]		✓			✓	✓	✓		
Hong2021 [33]	✓	✓	✓	✓	✓	✓	✓		
Lumini2021a [27]	✓	✓	✓		✓	✓	✓		
Lumini2021b [28]	✓	✓	✓		✓	✓	✓		
Nanni2021 [29]		✓			✓				
Thambawita2021 [36]		✓			✓		✓		
Tomar2021 [25]						✓			
Nanni2022a [37]		✓			✓				
Nanni2022b [30]	✓	✓	✓		✓	✓	✓		
Tran2022 [26]		✓							
Nanni2023 [38]		✓			✓				

Table 4. Ensembles for polyp segmentation: performance (Dice coefficient and IoU) reported by the authors, rounded to three significant digits. For papers that propose more than one ensemble, the table reports the figures for the best ensemble. Papers and datasets for which no scores are available do not appear in the table. Note that the results from different papers are, in general, not comparable because they were obtained with different testing protocols.

Paper	Kvasir		ClinDB		ColDB		ETIS		CVC-T		PolypGen		MediEval 2020	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Guo2019 [20]			0.967	0.983	0.962	0.980	0.970	0.985						
Kang2019 [31]					0.695		0.661							
Nguyen2019 [34]				0.891		0.896								
Shrestha2020 [32]	0.760	0.838											0.755	0.832
ThuHong2020 [35]					0.798	0.891	0.702	0.823						
Hong2021 [33]											0.571	0.619		
Lumini2021a [27]	0.820	0.885												
Lumini2021b [28]	0.825	0.888												
Nanni2021 [29]	0.872	0.919	0.886	0.931	0.701	0.776	0.663	0.743	0.831	0.901				
Thambawita2021 [36]											0.840			
Nanni2022a [37]	0.874	0.920	0.894	0.937	0.751	0.826	0.717	0.787	0.842	0.904				
Nanni2022b [30]	0.870	0.918	0.884	0.929	0.695	0.768	0.644	0.727	0.833	0.904				
Nanni2023 [38]	0.871	0.920	0.903	0.947	0.720	0.787	0.688	0.756	0.846	0.909				

3. Methods

In this section, we introduce and describe the methods adopted in this work, as well as the datasets and the functions used to assess the performance.

3.1. Structure of the Ensemble

As anticipated in Section 1, our ensemble is based on the HarDNet-MSEG [12], Polyp-PVT [13], and HSNet [14] network topologies. These models were selected for the following reasons: HarDNet-MSEG is a lightweight neural network that has shown very good performance in the polyp segmentation task; Polyp-PVT is one of the first attempts to use transformers in the field of polyp segmentation; HSNet is a mixed model including both CNN and transformers in the encoder. Our final ensemble includes multiple instances of the three models, trained in different ways as summarized in Section 3.4 and described in full in Section 4. In our preliminary experiments, we observed that, in these models, the last sigmoid layer followed by a normalization layer goes against the purpose of ensemble design because it pushes the scores to the extremes of the range, making the average or sum rule too similar to a voting rule.

A novelty in our network architecture is that we cut the normalization layer after the last sigmoid in HarDNet-MSEG, Polyp-PVT, and HSNet. In the original networks, before the output, each segmentation mask is normalized between $[0, 1]$: this implies that the networks always find a foreground object, but this assumption cannot be made in a real colonoscopy. Therefore, the reported results obtained using HarDNet-MSEG, Polyp-PVT, and HSNet are slightly different from the ones in the original papers.

Another significant difference is that, while in the original Polyp-PVT and HSNet topologies, the intermediate masks (two masks for Polyp-PVT and four for HSNet) are summed and then passed to the sigmoid, we passed each mask separately to the sigmoid and averaged the results. Consequently, our output is given by

$$\sum_{i=1}^n \text{sigmoid}(P_i) / n,$$

where P_i is an intermediate prediction mask and n is the number of the masks ($n = 2$ for Polyp-PVT, $n = 4$ for HSNet). In Figure 2, we compare the output of HSNet and the related four intermediate masks. It is clear that the output is sharper, so the sum rule is almost a voting rule if the original outputs of HSNet and Polyp-PVT are used in an ensemble.

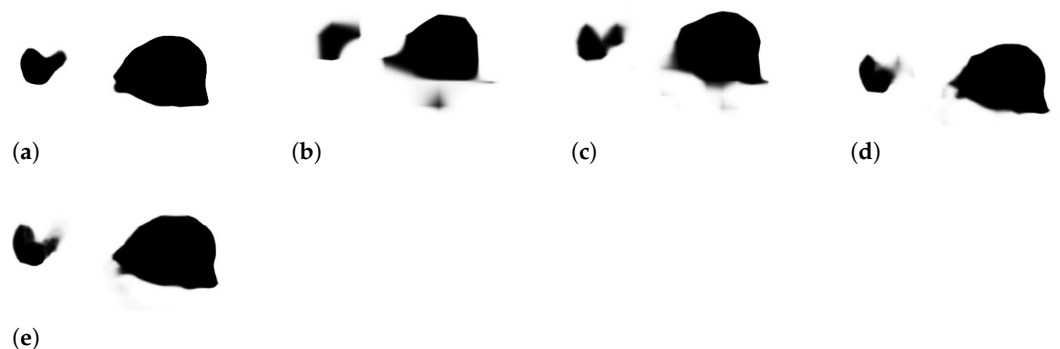


Figure 2. Masks in HSNet. (a) Final segmentation mask. (b–e) Intermediate masks after the sigmoid computation.

3.2. Performance Metrics and Loss Functions

In this section, we summarize the performance metrics and the loss functions adopted in this paper. For an exhaustive overview of image segmentation and loss functions, we point the interested reader to the recent survey [51].

We adopted the Dice coefficient [52] to measure the overlap between the predicted segmentation masks and the ground truth. This approach is widespread in semantic segmentation. The Dice coefficient is defined as

$$\text{Dice}(Y, T) = \frac{2 \times |Y \cap T|}{|Y| + |T|},$$

where Y is the predicted segmentation mask, T is the ground-truth mask, and the cardinality is the number of pixels.

Another well-known performance measure is the intersection over union (IoU), which was introduced in [53]:

$$\text{IoU}(Y, T) = \frac{|Y \cap T|}{|Y \cup T|}.$$

Hence, an IoU of 1 corresponds to a perfect prediction, that is a pixel-perfect overlap between the predicted segmentation mask and the ground truth. The corresponding loss function is defined as

$$L'_{IoU} = 1 - \text{IoU}.$$

This could be an issue when dealing with imbalanced datasets. Therefore, as suggested in [54], we used the weighted intersection over union (wIoU) instead of the standard IoU. The corresponding loss function is

$$L'_{wIoU} = 1 - \frac{1 + \sum_{i=1}^N \sum_{k=1}^K w_{ik} T_{ik} * Y_{ik}}{1 + \sum_{i=1}^N \sum_{k=1}^K w_{ik} (T_{ik} + Y_{ik} - T_{ik} * Y_{ik})},$$

where N is the number of pixels, K is the number of classes, and w_{ik} is the weight given to the i -th pixel of the image for the class k . These weights were computed as before. T_{ik} and Y_{ik} are, respectively, the ground truth value and the prediction value for the i -th pixel belonging to the class k . We added 1 to both the numerator and the denominator to prevent undefined divisions.

The cross-entropy (CE) loss function provides us with a measure of the difference between two probability distributions. The goal is to minimize this difference, and in doing so, it has no bias between small or large regions. This could be an issue when dealing with imbalanced datasets. Hence, the weighted CE loss was introduced, and it resulted in well-balanced classifiers for imbalanced scenarios [55]. The formula for the weighted binary CE loss is

$$L_{wBCE} = - \sum_{i=1}^N \sum_{k=1}^K w_{ik} T_{ik} * \log(P_{ik}),$$

where N is the number of pixels, K is the number of classes, and w_{ik} is the weight given to the i -th pixel of the image for the class k . These weights were computed by using an average pooling over the mask with a kernel of size 31×31 and a stride of 1 to consider also non-maximal activations. T_{ik} is the true value for the i -th pixel, and it can be equal to either 0 or 1. It is 1 if the i -th pixel belongs to the class k , 0 otherwise. P_{ik} is the probability that the i -th pixel belongs to the class k obtained by using the sigmoid activation function. For P , we used the softmax activation function, which returns probabilities.

Based on the intuition in [12], the wIoU loss and the weighted binary CE loss were considered together (structure loss function):

$$L'_{STR} = L_{wIoU} + L_{wBCE}.$$

In our experiments, the structure loss function was used to train all the networks except DeepLabv3+.

3.3. Datasets and Data Augmentations

Polyp segmentation from colonoscopy images is a challenging task that requires two-class discrimination between polyp pixels and the low-contrast colon background. We present experimental results on five datasets for polyp segmentation:

- The Kvasir-SEG [56] dataset (“Kvasir”) contains medical images that have been labeled and verified by doctors. The images depict different parts of the digestive system and show both healthy and diseased tissue. The dataset includes images at different resolutions (from 720×576 up to 1920×1072 pixels) and is organized into folders based on the content of the images. Some of the images also include a small picture-in-picture showing the position of the endoscope inside the body.
- CVC-ColonDB [7] (“ColDB”) is a dataset of 300 images that aims to include a wide range of appearances for polyps. The goal is to provide as much diversity as possible in the dataset.
- CVC-T (sometimes called “Endo”) is the test set of a larger dataset named CVC-EndoSceneStill [47].
- The ETIS-Larib [4] dataset (“ETIS”) contains 196 colonoscopy images. Two frames from the dataset and their corresponding ground truth masks are shown in Figure 1.
- CVC-ClinicDB [39] (“ClinDB”) contains 612 images from 31 videos of colonoscopy procedures. The images have been manually labeled by experts to identify the regions covered by the polyps, and ground truth information is also provided for light reflections. The images are 576×768 pixels in size.

Our training set is made of 1450 images taken from the largest datasets, namely 900 images from Kvasir and 550 images from ClinDB. The remaining images (100 from Kvasir, 62 from ClinDB, plus all the images from ColDB, CVC-T, and ETIS) were used for the test set. As pointed out in Section 2, this experimental protocol was proposed in [10] and has been followed by many, including the researchers behind the three base networks we considered (see Section 3.1). Having a small training set is a common challenge in deep learning and can often lead to overfitting, i.e., the model memorizes the training data rather than learning generalizable patterns. In this work, we adopted two common techniques for addressing this issue: fine-tuning and data augmentation. Fine-tuning was used since all the models involved in our experiments were pre-trained on a large dataset. Data augmentation is based on two different strategies, which aim to increase the effective size of the training set and provide the model with more examples to learn from:

- “DA1”: a basic strategy that includes two image flips (up/down and left/right) and a 90-degree counterclockwise rotation; therefore, three synthetic images were created for each original image.
- “DA2”: a sophisticated strategy that creates synthetic images in 13 different ways, including the application of motion blur and shadows to the original images.

The two strategies were introduced in [37]; we refer the interested reader to this work for a thorough description. Figure 3 highlights the images provided by the data augmentation process for a sample polyp image.

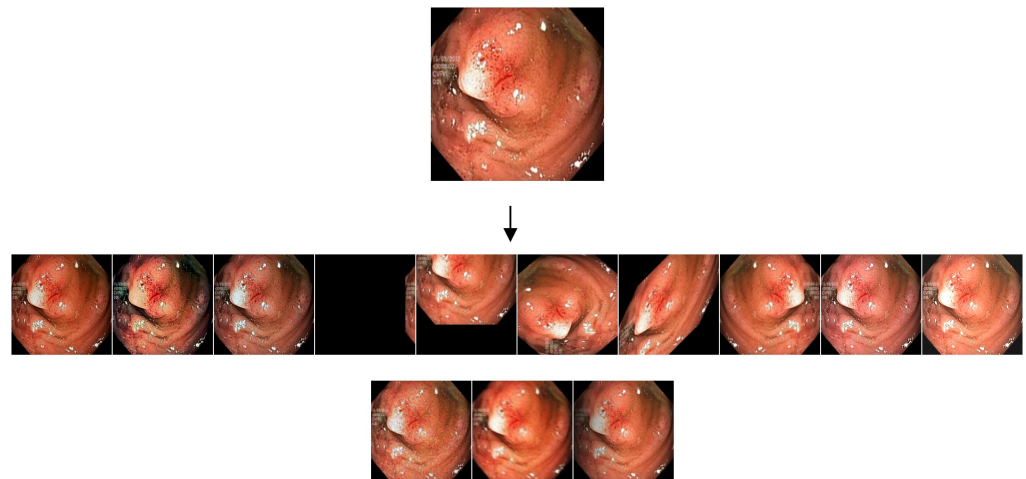


Figure 3. An example of images obtained through our data augmentation process. **(Top)** Original polyp image. **(Bottom)** Synthetic images created from the original image by data augmentation.

3.4. Overview of the Experiments

In our experiments, we considered different ensembles built with the base networks HarDNet-MSEG, Polyp-PVT, and HSNet (see Section 3.1). Polyp-PVT and HSNet are based on transformers, introduced recently [57] in the domain of polyp segmentation. Figure 4 shows the overall structure of the network for the most-complex ensemble we investigated in the experimental section (Section 4, Ensemble Ens3). The networks were trained for 100 epochs with a batch size of 20 for HarDNet-MSEG and 8 for Polyp-PVT and HSNet. To avoid any overfitting, we used the default value for the number of epochs. The networks were always trained using the structure loss function (see Section 3.2) except for DeepLabv3+, and their outputs were combined with the sum rule or the weighted sum rule depending on the experiment. We used resized training images of size 352×352 . During the test phase, masks were obtained by resizing the images and were subsequently scaled back to the original dimensions to evaluate the performance of the model.

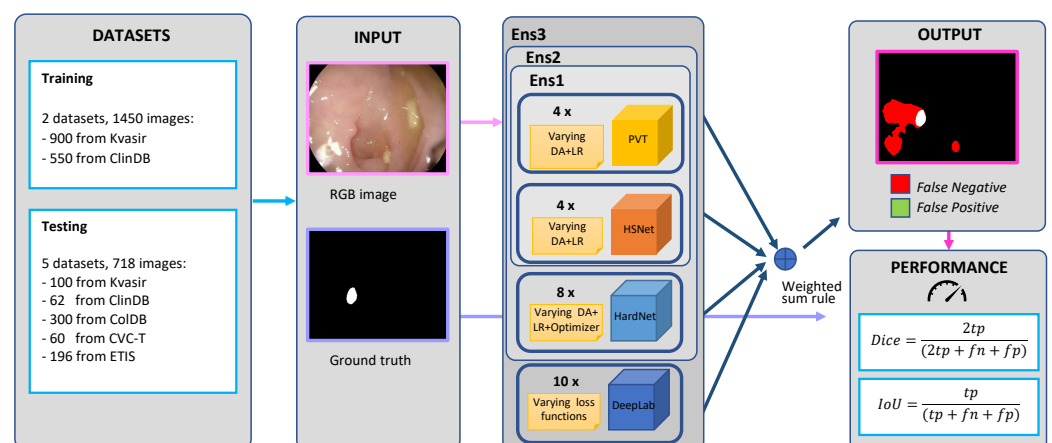


Figure 4. Overall structure of the proposed method, including the training and testing procedures.

As optimization methods, we experimented with both Adam and stochastic gradient descent (SGD) for HarDNet-MSEG, while we adopted AdamW for Polyp-PVT and HSNet, like in the original papers.

We trained the networks with the two data augmentation techniques described in Section 3.3. We also experimented with training two identical networks with the two techniques and combining their outputs.

Finally, the following two learning rates were considered in the experiments:

- 10^{-4} (learning rate “a”);
- 5×10^{-5} decaying to 5×10^{-6} after 30 epochs (learning rate “b”).

We also trained some networks twice with the two learning rates and, as we did with the data augmentations, combined their outputs.

4. Experimental Results

In this section, we report on the experimental analysis carried out to assess the proposed ensemble strategy.

4.1. Ablation Studies

Ablation in machine learning refers to removing parts of a machine learning model to measure their impact on the performance. In this work, we employed an ensemble approach by combining the predictions of the three models to improve the overall performance of our system. To evaluate the contributions of the different components in our ensemble, we conducted an ablation study by separately analyzing the potential of each model. The results are summarized in Tables 5–7, respectively. Note that the performance of the ensembles using multiple models will be described later, in Section 4.2. The Dice coefficient was used to measure the performance on the five test sets defined in Section 3.3. In the column DA, we specify which data augmentation approaches (see Section 3.3) were considered. Where the column states “1 + 2”, we trained the network twice with Data Augmentations 1 and 2. In the column LR, we specify which learning rates (see Section 3.4) were used during training. Where the column states “a + b”, we trained the network twice with Learning Rates a and b. In Table 5, the additional column OPT specifies which optimization methods (see Section 3.4) were adopted:

- “SGD”: stochastic gradient descent.
- “Adam”: Adam.
- “SGD + Adam”: both.

If a network was trained multiple times with different choices, the outputs were combined with the sum rule, and the final model was, therefore, an ensemble. Consequently, the last three rows in the table represent ensembles of 4, 4, and 8 HarDNet-MSEG networks, respectively.

Table 5. Dice coefficient for the (ensemble) networks built on HarDNet-MSEG. The best performance figures are marked in bold.

OPT	DA	LR	Kvasir	ClinDB	ColDB	ETIS	CVC-T	Avg
SGD	1	a	0.893	0.875	0.745	0.667	0.882	0.812
		b	0.862	0.794	0.705	0.628	0.873	0.772
SGD	2	a	0.914	0.944	0.747	0.727	0.901	0.847
		b	0.871	0.875	0.702	0.650	0.885	0.797
Adam	1	a	0.906	0.924	0.751	0.716	0.903	0.840
		b	0.910	0.910	0.748	0.702	0.884	0.831
Adam	2	a	0.896	0.927	0.778	0.774	0.893	0.854
		b	0.895	0.916	0.758	0.716	0.885	0.834
SGD + Adam	1 + 2	a	0.918	0.947	0.778	0.756	0.909	0.862
		b	0.907	0.928	0.775	0.770	0.899	0.856
		a + b	0.915	0.931	0.785	0.781	0.904	0.863

The conclusions we can draw from the results in Table 5 are that, on average, Learning Rate a provided better results than Learning Rate b. Learning Rate b appeared to perform poorly when coupled with SGD. The best average performance was provided by the ensemble that adopts both different data augmentation strategies and different learning rates (last row of Table 5); however, the Dice coefficient was only marginally better than that of the ensemble trained with Learning Rate a alone (Row 9).

In Tables 6 and 7, the additional column SM specifies whether we obtained the final segmentation mask as in the original HSNNet and Polyp-PVT networks (SM = “No”) or with the novel approach we proposed in Section 3.1 (SM = “Yes”).

Table 6. Dice coefficient for the (ensemble) networks built on Polyp-PVT. The best performance figures are marked in bold.

DA	LR	SM	Kvasir	ClinDB	CoIDB	ETIS	CVC-T	Avg
1	a	No	0.911	0.926	0.788	0.773	0.871	0.854
1	b	No	0.924	0.924	0.793	0.800	0.877	0.864
2	a	No	0.910	0.923	0.804	0.749	0.891	0.855
2	b	No	0.917	0.921	0.794	0.763	0.891	0.857
1 + 2	a	No	0.918	0.926	0.803	0.755	0.873	0.855
1 + 2	a	Yes	0.919	0.930	0.809	0.765	0.884	0.861
1 + 2	b	No	0.931	0.920	0.792	0.776	0.876	0.859
1 + 2	b	Yes	0.931	0.921	0.798	0.791	0.882	0.865
1 + 2	a + b	No	0.926	0.932	0.821	0.800	0.891	0.874
1 + 2	a + b	Yes	0.926	0.933	0.824	0.808	0.895	0.877

Table 7. Dice coefficient for the (ensemble) networks built on HSNet. The best performance figures are marked in bold.

DA	LR	SM	Kvasir	ClinDB	CoIDB	ETIS	CVC-T	Avg
1	a	No	0.923	0.921	0.789	0.733	0.898	0.853
1	b	No	0.930	0.934	0.821	0.783	0.901	0.873
2	a	No	0.909	0.944	0.806	0.750	0.901	0.862
2	b	No	0.913	0.947	0.816	0.783	0.903	0.872
1 + 2	a	No	0.923	0.925	0.794	0.703	0.891	0.847
1 + 2	a	Yes	0.922	0.929	0.808	0.746	0.903	0.862
1 + 2	b	No	0.931	0.938	0.821	0.775	0.896	0.872
1 + 2	b	Yes	0.929	0.943	0.822	0.783	0.903	0.876
1 + 2	a + b	No	0.928	0.943	0.829	0.779	0.902	0.876
1 + 2	a + b	Yes	0.928	0.945	0.828	0.791	0.905	0.879

The conclusions we can draw from the results reported in Tables 6 and 7 are that, on average, Learning Rate b performed better than Learning Rate a when coupled with HSNet and slightly better when coupled with Polyp-PVT. For both Polyp-PVT and HSNet, the best average performance was obtained by the ensemble obtained by varying the data augmentation and the learning rate.

Using the proposed approach to obtain the segmentation masks allowed us to increase the performance of the Polyp-PVT/HSNet ensembles. Some inference masks obtained using HSNet ensembles (HSNet a + b) with and without smoothing are reported in Figure 5, where false positive pixels are highlighted in green, while the false negatives are in red. They demonstrate that our ensemble model with smoothing produced better boundary results and made more accurate predictions than the models without smoothing.

In Table 8, we report the output values before the sigmoid layer for HarDNet-MSEG and the intermediate masks of Polyp-PVT and HSNet. All the networks were trained with Data Augmentation 1 and Learning Rate a. In the last two columns, the non-saturation rate of the resulting masks is reported using our averaging rule, compared to the original sum rule followed by normalization. Non-saturation is measured as the percentage p of pixels such that $abs(p) < 6.9$ (since $sigmoid(6.9) \approx 0.999$). It can be observed that these output values are all very close to saturating the sigmoid function. Summing them together would exacerbate the situation by producing an almost binary output. As shown in the last two columns, our approach has a higher rate of non-saturating pixels in the sigmoid function. Our approach averages the intermediate masks and produces a smoother output, which means that it maintains more information. This information can be visually appreciated in Figure 5. In particular, it can be noticed that the final segmentation mask of HSNet (Figure 5a) is very sharp, indicating an almost binary output, while the four intermediate masks P_1 – P_4 have more blurred edges.

Table 8. Output values before the sigmoid layer. P_i is the i -th intermediate prediction mask. Avg_{min} is the average of the minimum values (for each image) before the sigmoid. Avg_{max} is the average of the maximum values (for each image) before the sigmoid. Max_{min} is the max of the minimum values. Min_{max} is the min of the maximum values. Sat_{Avg} is the rate of non-saturated pixels for fusion by the averaging rule. Sat_{Sum} is the rate of non-saturated pixels for fusion by the sum rule. Fusion is among all P_i s returned by a model.

Method		Min_{max}	Avg_{max}	Avg_{min}	Max_{min}	Sat_{Avg}	Sat_{Sum}
HarDNet-MSEG	P	2.51	71.84	−6.53	−11.45	2.57%	2.57%
Polyp-PVT	P_1	26.22	70.15	−21.06	−30.04	1.62%	1.09%
	P_2	19.53	38.55	−20.61	−29.47		
HSNet	P_1	24.21	67.81	−38.49	−69.06	2.92%	0.85%
	P_2	33.19	87.04	−42.63	−84.26		
	P_3	31.31	122.01	−63.13	−123.45		
	P_4	36.47	160.11	−95.23	−165.40		

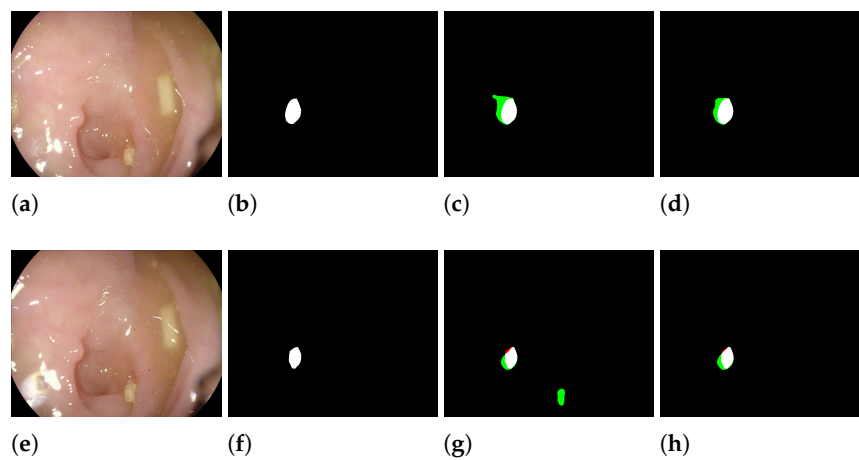


Figure 5. Masks in HSNet ensemble with and without smoothing. (a,e) Original image, (b,f) ground truth, (c,g) HSNet ensemble without smoothing, and (d,h) HSNet ensemble with smoothing. False positive pixels are green. False negative pixels are red.

4.2. Proposed Ensembles and Comparison with State-of-the-Art Models

In this section, we compared the performance of the three ensembles proposed in this work with twenty state-of-the-art methods for polyp segmentation (Table 9). In our final proposed ensembles, the models analyzed in Section 4.1 were combined with the weighted sum rule. Each method was weighted so that its weight in the fusion is equal to the other methods. We report the performance of the following ensembles:

- “Ens1”: an ensemble of 4 Polyp-PVT networks with the segmentation masks obtained with our approach and trained with all possible combinations of Data Augmentations 1 and 2 and Learning Rates a and b (Table 6, last row), plus 4 HSNet networks with the segmentation masks obtained with our approach and trained with all possible combinations of Data Augmentations 1 and 2 and Learning Rates a and b (Table 7, last row).
- “Ens2”: like Ens1, plus eight HarDNet-MSEG networks trained with all possible combinations of the SGD and Adam optimizers, Data Augmentations 1 and 2, and Learning Rates a and b (Table 5, last row).
- “Ens3”: like Ens2, plus the best ensemble introduced in [37] based on DeepLabv3+.

From the results reported in Table 9, it can be seen that the proposed ensembles beat the state-of-the-art on the ColDB and ETIS datasets, which can be regarded as the most-challenging since the corresponding Dice scores were consistently lower not only for the ensemble components and ensembles (Tables 5–7 and 9, Rows 1–3), but also for competing solutions in the literature (Table 9, Rows 4–23). We remark that this result was

obtained without training the ensembles on ColDB and ETIS. Most importantly, all three proposed ensembles performed better than state-of-the-art solutions when we averaged across all datasets (Table 9, last two columns). Even the simplest of our ensembles, that is Ens1, beat the state-of-the-art on average. The conclusion we can draw is that the proposed ensembles were strong performers with all the datasets: even when they were not the best, they were near the top. This is a benefit of the ensemble strategy. A change in performance with the dataset remains unavoidable: images in different datasets were obtained with different acquisition instruments, under different conditions, and created by experts with different opinions on what is important to include in the dataset. For instance, the authors of ColDB deemed it important to exclude similar frames from the dataset and, in general, to maximize the variability between polyps in different images. Again, this makes ColDB a challenging dataset. Of course, the three ensembles performed significantly better than their baseline networks (Table 9, Rows 4–7), including the recent ensemble introduced in [37]. The usefulness of transformers emerges clearly from the performance difference between Ens1 (transformers) and HarDNet-MSEG (best CNN tested). Polyp-PVT and HSNet were the best transformer-based methods; their fusion in Ens1 allowed for clearly better performance than that obtained by any of the two methods alone, as seen by comparing Ens1 with Polyp-PVT and HSNet. On the other hand, adding CNN-based methods to Ens1, as was done in Ens2 and Ens3, only led to marginal improvements.

Table 9. Performance comparison between the proposed ensembles and recent models available in the literature. The best performance figures are marked in bold. Underlined figures point out where our ensembles beat, or were on par with, the state-of-the-art.

Method	Kvasir		ClinDB		ColDB		ETIS		CVC-T		Avg	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
Ens1	0.886	0.930	0.892	0.936	0.764	0.839	0.738	<u>0.812</u>	0.841	0.904	<u>0.824</u>	0.884
Ens2	0.883	0.927	0.894	0.938	<u>0.759</u>	<u>0.832</u>	<u>0.750</u>	<u>0.822</u>	0.841	0.904	<u>0.825</u>	0.885
Ens3	0.882	0.928	0.894	0.937	<u>0.752</u>	0.824	0.760	0.830	0.840	0.905	0.826	0.885
HarDNet-MSEG [12]	0.857	0.912	0.882	0.932	<u>0.66</u>	0.731	0.613	0.677	0.821	0.887	0.767	0.828
Polyp-PVT [13]	0.864	0.917	0.889	0.937	0.727	0.808	0.706	0.787	0.833	0.9	0.804	0.869
HSNet [14]	0.877	0.926	0.905	0.948	0.735	0.81	0.734	0.808	0.839	0.903	0.818	0.879
FH(2) + 2 × PVT(2) [37]	0.874	0.920	0.894	0.937	0.751	0.826	0.717	0.787	0.842	0.904	0.816	0.875
PraNet (from [12])	0.84	0.898	0.849	0.899	0.64	0.709	0.567	0.628	0.797	0.871	0.739	0.801
SFA (from [12])	0.611	0.723	0.607	0.7	0.347	0.469	0.217	0.297	0.329	0.467	0.422	0.531
U-Net++ (from [12])	0.743	0.821	0.729	0.794	0.41	0.483	0.344	0.401	0.624	0.707	0.57	0.641
U-Net (from [12])	0.746	0.818	0.755	0.823	0.444	0.512	0.335	0.398	0.627	0.71	0.581	0.652
Eloss101-Mix + FH [38]	0.871	0.920	0.903	0.947	0.720	0.787	0.688	0.756	0.846	0.909	0.806	0.864
MIA-Net [58]	0.876	0.926	0.899	0.942	0.739	0.816	0.725	0.8	0.835	0.9	0.815	0.877
P2T [59]	0.849	0.905	0.873	0.923	0.68	0.761	0.631	0.7	0.805	0.879	0.768	0.834
DBMF [60]	0.886	0.932	0.886	0.933	0.73	0.803	0.711	0.79	0.859	0.919	0.814	0.875
SETR [61]	0.854	0.911	0.885	0.934	0.69	0.773	0.646	0.726	0.814	0.889	0.778	0.847
TransUnet [62]	0.857	0.913	0.887	0.935	0.699	0.781	0.66	0.731	0.824	0.893	0.785	0.851
TransFuse [11]	0.87	0.92	0.897	0.942	0.706	0.781	0.663	0.737	0.826	0.894	0.792	0.855
UACANet [63]	0.859	0.912	0.88	0.926	0.678	0.751	0.678	0.751	0.849	0.91	0.789	0.85
SANet [64]	0.847	0.904	0.859	0.916	0.67	0.753	0.654	0.75	0.815	0.888	0.769	0.842
MSNet [65]	0.862	0.907	0.879	0.921	0.678	0.755	0.664	0.719	0.807	0.869	0.778	0.834
SwinE-Net [66]	0.87	0.92	0.892	0.938	0.725	0.804	0.687	0.758	0.842	0.906	0.803	0.865
AMNet [67]	0.865	0.912	0.888	0.936	0.69	0.762	0.679	0.756	-	-	-	-

5. Conclusions

In this work, we provided a review of the literature on deep learning ensembles for polyp segmentation and demonstrated the advantages of tackling semantic segmentation with ensembles of convolutional and transformer neural networks. The main idea behind ensembling is to combine the predictions of multiple models to improve the overall performance. We introduced an effective way of doing this by averaging intermediate predictions in a new fashion. This can help to smooth out the contribution of any specific model, besides reducing the impact of overfitting to a particular dataset.

We plan to generalize our results to other application domains. For this reason, many datasets will be used in the future to corroborate the conclusions reported here, namely to prove that:

- A fusion of different convolutional and transformer topologies can achieve state-of-the-art performance;

- Applying different approaches to the learning rate strategy is a feasible method to build a set of segmentation networks;
- A better way to add the transformers (Polyp-PVT and HSNNet) in an ensemble is to use the proposed approach for creating the final segmentation mask.

Furthermore, we plan to test our model with different distillation techniques and pruning approaches to adapt this technique to low-cost hardware. This will allow us to extend the usefulness of our model to situations where the available computational power is restricted or constrained.

Author Contributions: Conceptualization, L.N. and C.F.; software, L.N. and A.L. (Alessandra Lumini); writing—original draft preparation, C.F., A.L. (Andrea Loreggia) and L.N.; writing—review and editing, C.F., A.L. (Andrea Loreggia), L.N. and A.L. (Alessandra Lumini). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the resources required to replicate our experiments are available at <https://github.com/LorisNanni> (accessed on 8 May 2023).

Acknowledgments: We would like to acknowledge the support that NVIDIA provided us through the GPU Grant Program. We used a donated TitanX GPU to train the neural networks discussed in this work.

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

References

1. Siegel, R.L.; Miller, K.D.; Sauer, A.G.; Fedewa, S.A.; Butterly, L.F.; Anderson, J.C.; Cercek, A.; Smith, R.A.; Jemal, A. Colorectal cancer statistics, 2020. *CA Cancer J. Clin.* **2020**, *70*, 145–164. [[CrossRef](#)] [[PubMed](#)]
2. Valderrama-Treviño, A.; Hazzel, E.; Flores, C.; Herrera, M. Colorectal cancer: A review. *Artic. Int. J. Res. Med. Sci.* **2017**, *5*, 4667–4676. [[CrossRef](#)]
3. Wieszczy, P.; Regula, J.; Kaminski, M.F. Adenoma detection rate and risk of colorectal cancer. *Best Pract. Res. Clin. Gastroenterol.* **2017**, *31*, 441–446. [[CrossRef](#)]
4. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *Int. J. Comput. Assist. Radiol. Surg.* **2014**, *9*, 283–293. [[CrossRef](#)]
5. Mamonov, A.V.; Figueiredo, I.N.; Figueiredo, P.N.; Richard Tsai, Y.H. Automated polyp detection in colon capsule endoscopy. *IEEE Trans. Med. Imaging* **2014**, *33*, 1488–1502. [[CrossRef](#)] [[PubMed](#)]
6. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Trans. Med. Imaging* **2016**, *35*, 630–644. [[CrossRef](#)] [[PubMed](#)]
7. Bernal, J.; Sánchez, J.; Vilariño, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognit.* **2012**, *45*, 3166–3182. [[CrossRef](#)]
8. Minaee, S.; Boykov, Y.; Porikli, F.; Plaza, A.; Kehtarnavaz, N.; Terzopoulos, D. Image Segmentation Using Deep Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 3523–3542. [[CrossRef](#)]
9. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; Lange, T.D.; Halvorsen, P.; D. Johansen, H. ResUNet++: An Advanced Architecture for Medical Image Segmentation. In Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–230. [[CrossRef](#)]
10. Fan, D.P.; Ji, G.P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2020*; Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L., Eds.; Springer: Cham, Switzerland, 2020; pp. 263–273. [[CrossRef](#)]
11. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., Eds.; Springer: Cham, Switzerland, 2021; pp. 14–24. [[CrossRef](#)]
12. Huang, C.H.; Wu, H.Y.; Lin, Y.L. HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS. *arXiv* **2021**, arXiv:2101.07172.
13. Dong, B.; Wang, W.; Fan, D.P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers. *arXiv* **2021**, arXiv:2108.06932.
14. Zhang, W.; Fu, C.; Zheng, Y.; Zhang, F.; Zhao, Y.; Sham, C.W. HSNNet: A hybrid semantic network for polyp segmentation. *Comput. Biol. Med.* **2022**, *150*, 106173. [[CrossRef](#)] [[PubMed](#)]
15. Cornelio, C.; Donini, M.; Loreggia, A.; Pini, M.S.; Rossi, F. Voting with random classifiers (VORACE): Theoretical and experimental analysis. *Auton. Agent* **2021**, *35*, 2. [[CrossRef](#)]

16. Pacal, I.; Karaboga, D.; Basturk, A.; Akay, B.; Nalbantoglu, U. A comprehensive review of deep learning in colon cancer. *Comput. Biol. Med.* **2020**, *126*, 104003. [[CrossRef](#)] [[PubMed](#)]
17. Zhang, W.; Zhuang, P.; Sun, H.H.; Li, G.; Kwong, S.; Li, C. Underwater Image Enhancement via Minimal Color Loss and Locally Adaptive Contrast Enhancement. *IEEE Trans. Image Process.* **2022**, *31*, 3997–4010. [[CrossRef](#)]
18. Nisha, J.; P. Gopi, V.; Palanisamy, P. Automated colorectal polyp detection based on image enhancement and dual-path CNN architecture. *Biomed. Signal Process. Control* **2022**, *73*, 103465. [[CrossRef](#)]
19. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer: Cham, Switzerland, 2015; pp. 234–241.
20. Guo, X.; Zhang, N.; Guo, J.; Zhang, H.; Hao, Y.; Hang, J. Automated polyp segmentation for colonoscopy images: A method based on convolutional neural networks and ensemble learning. *Med. Phys.* **2019**, *46*, 5666–5676. [[CrossRef](#)]
21. Ji, G.P.; Xiao, G.; Chou, Y.C.; Fan, D.P.; Zhao, K.; Chen, G.; Van Gool, L. Video polyp segmentation: A deep learning perspective. *Mach. Intell. Res.* **2022**, *19*, 531–549. [[CrossRef](#)]
22. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
23. Singh, N.K.; Raza, K. Medical image generation using generative adversarial networks: A review. *Health Informatics: A Computational Perspective in Healthcare*; Springer: Singapore, 2021; pp. 77–96.
24. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Front. Comput. Sci.* **2020**, *14*, 241–258. [[CrossRef](#)]
25. Tomar, N.K.; Ibtehaz, N.; Jha, D.; Halvorsen, P.; Ali, S. Improving Generalizability in Polyp Segmentation using Ensemble Convolutional Neural Network. In Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) Co-Located with the 18th IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France, 13 April 2021; Volume 2886; pp. 49–58.
26. Tran, T.N.; Isensee, F.; Krämer, L.; Yamlahi, A.; Adler, T.; Godau, P.; Tizabi, M.; Maier-Hein, L. Heterogeneous Model Ensemble For Automatic Polyp Segmentation In Endoscopic Video Sequences. In Proceedings of the 4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2022) Co-Located with the 19th IEEE International Symposium on Biomedical Imaging (ISBI 2022), Kolkata, India, 28–31 March 2022; Volume 3148, pp. 20–24.
27. Lumini, A.; Nanni, L.; Maguolo, G. Deep ensembles based on Stochastic Activation Selection for Polyp Segmentation. *arXiv* **2021**, arXiv:2104.00850.
28. Lumini, A.; Nanni, L.; Maguolo, G. Deep Ensembles Based on Stochastic Activations for Semantic Segmentation. *Signals* **2021**, *2*, 820–833. [[CrossRef](#)]
29. Nanni, L.; Cuza, D.; Lumini, A.; Loreggia, A.; Brahnam, S. Deep ensembles in bioimage segmentation. *arXiv* **2021**, arXiv:2112.12955.
30. Nanni, L.; Cuza, D.; Lumini, A.; Brahnam, S. Data augmentation for deep ensembles in polyp segmentation. In *Computational Intelligence Based Solutions for Vision Systems*; IOP Publishing: Bristol, UK, 2022; pp. 8-1–8-22. [[CrossRef](#)]
31. Kang, J.; Gwak, J. Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images. *IEEE Access* **2019**, *7*, 26440–26447. Erratum in *IEEE Access* **2020**, *8*, 100010–100012. [[CrossRef](#)]
32. Shrestha, S.; Khanal, B.; Ali, S. Ensemble U-Net Model for Efficient Polyp Segmentation. In Proceedings of the MediaEval 2020 Workshop, Online, 14–15 December 2020; Volume 2882.
33. Hong, A.; Lee, G.; Lee, H.; Seo, J.; Yeo, D. Deep Learning Model Generalization with Ensemble in Endoscopic Images. In Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) Co-Located with the 18th IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France, 13 April 2021; Volume 2886; pp. 80–89.
34. Nguyen, N.Q.; Lee, S.W. Robust Boundary Segmentation in Medical Images Using a Consecutive Deep Encoder-Decoder Network. *IEEE Access* **2019**, *7*, 33795–33808. [[CrossRef](#)]
35. Thu Hong, L.T.; Chi Thanh, N.; Long, T.Q. Polyp Segmentation in Colonoscopy Images Using Ensembles of U-Nets with EfficientNet and Asymmetric Similarity Loss Function. In Proceedings of the 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, 14–15 October 2020; pp. 1–6. [[CrossRef](#)]
36. Thambawita, V.; Hicks, S.; Halvorsen, P.; Riegler, M. DivergentNets: Medical Image Segmentation by Network Ensemble. In Proceedings of the 3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021) Co-Located with the 18th IEEE International Symposium on Biomedical Imaging (ISBI 2021), Nice, France, 13 April 2021; Volume 2886, pp. 27–38.
37. Nanni, L.; Lumini, A.; Loreggia, A.; Formaggio, A.; Cuza, D. An Empirical Study on Ensemble of Segmentation Approaches. *Signals* **2022**, *3*, 341–358. [[CrossRef](#)]
38. Nanni, L.; Cuza, D.; Lumini, A.; Loreggia, A.; Brahman, S., Polyp Segmentation with Deep Ensembles and Data Augmentation. In *Artificial Intelligence and Machine Learning for Healthcare*; Vol. 1: Image and Data Analytics; Springer International Publishing: Cham, Switzerland, 2023; pp. 133–153. [[CrossRef](#)]
39. Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; Gil, D.; Rodríguez, C.; Vilariño, F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Comput. Med. Imaging Graph.* **2015**, *43*, 99–111. [[CrossRef](#)] [[PubMed](#)]
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]

41. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 833–851. [[CrossRef](#)]
42. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR: Cambridge, MA, USA, Volume 97, pp. 6105–6114.
43. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Halvorsen, P.; de Lange, T.; Johansen, D.; Johansen, H.D. Kvasir-SEG: A Segmented Polyp Dataset. In *MultiMedia Modeling*; Ro, Y.M., Cheng, W.H., Kim, J., Chu, W.T., Cui, P., Choi, J.W., Hu, M.C., De Neve, W., Eds.; Springer: Cham, Switzerland, 2020; pp. 451–462. [[CrossRef](#)]
44. Hosseinzadeh Kassani, S.; Hosseinzadeh Kassani, P.; Wesolowski, M.J.; Schneider, K.A.; Deters, R. Automatic Polyp Segmentation Using Convolutional Neural Networks. In *Advances in Artificial Intelligence*; Goutte, C., Zhu, X., Eds.; Springer: Cham, Switzerland, 2020; pp. 290–301. [[CrossRef](#)]
45. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J.M.R., Bradley, A., Papa, J.P., Belagiannis, V., et al., Eds.; Springer: Cham, Switzerland, 2018; pp. 3–11. [[CrossRef](#)]
46. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017*; pp. 936–944. [[CrossRef](#)]
47. Vázquez, D.; Bernal, J.; Sánchez, F.J.; Fernández-Esparrach, G.; López, A.M.; Romero, A.; Drozdal, M.; Courville, A. A benchmark for endoluminal scene segmentation of colonoscopy images. *J. Healthc. Eng.* **2017**, *2017*, 4037190. [[CrossRef](#)] [[PubMed](#)]
48. Baheti, B.; Innani, S.; Gajre, S.; Talbar, S. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020*; pp. 1473–1481. [[CrossRef](#)]
49. Isensee, F.; Jaeger, P.F.; Kohl, S.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [[CrossRef](#)]
50. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
51. Jadon, S. A survey of loss functions for semantic segmentation. *arXiv* **2020**, arXiv:2006.14822.
52. Sudre, C.H.; Li, W.; Vercauteren, T.; Ourselin, S.; Cardoso, M.J. Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations. *arXiv* **2017**, arXiv:1707.03237.
53. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*; Springer: Cham, Switzerland, 2016; pp. 234–244.
54. Cho, Y.J. Weighted Intersection over Union (wIoU): A New Evaluation Metric for Image Segmentation. *arXiv* **2021**, arXiv:2107.09858.
55. Aurelio, Y.S.; de Almeida, G.M.; de Castro, C.L.; Braga, A.P. Learning from imbalanced datasets with weighted cross-entropy function. *Neural Process. Lett.* **2019**, *50*, 1937–1949. [[CrossRef](#)]
56. Pogorelov, K.; Randel, K.R.; Griwodz, C.; Eskeland, S.L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D.T.; Lux, M.; Schmidt, P.T.; et al. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan, 20–23 June 2017*; pp. 164–169.
57. Shen, T.; Xu, H. Medical image segmentation based on Transformer and HardNet structures. *IEEE Access* **2023**, *11*, 16621–16630. [[CrossRef](#)]
58. Li, W.; Zhao, Y.; Li, F.; Wang, L. MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation. *Knowl.-Based Syst.* **2022**, *247*, 108824. [[CrossRef](#)]
59. Wu, Y.H.; Liu, Y.; Zhan, X.; Cheng, M.M. P2T: Pyramid Pooling Transformer for Scene Understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *early access*. [[CrossRef](#)]
60. Liu, F.; Hua, Z.; Li, J.; Fan, L. DBMF: Dual Branch Multiscale Feature Fusion Network for polyp segmentation. *Comput. Biol. Med.* **2022**, *151*, 106304. [[CrossRef](#)]
61. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.; et al. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021*; pp. 6877–6886. [[CrossRef](#)]
62. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. *arXiv* **2021**, arXiv:2102.04306. <https://doi.org/10.48550/ARXIV.2102.04306>.
63. Kim, T.; Lee, H.; Kim, D. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia—MM '21, Virtual, China, 20–24 October 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 2167–2175. [[CrossRef](#)]
64. Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S.K.; Cui, S. Shallow Attention Network for Polyp Segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Cham, Switzerland, 2021; Volume 12901. [[CrossRef](#)]
65. Zhao, X.; Zhang, L.; Lu, H. Automatic Polyp Segmentation via Multi-scale Subtraction Network. *arXiv* **2021**, arXiv:2108.05082.

66. Park, K.B.; Lee, J.Y. SwinE-Net: Hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *J. Comput. Des. Eng.* **2022**, *9*, 616–632. [[CrossRef](#)]
67. Song, P.; Li, J.; Fan, H. Attention based multi-scale parallel network for polyp segmentation. *Comput. Biol. Med.* **2022**, *146*, 105476. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.