


## Article

# Ensembles of Convolutional Neural Networks and Transformers for Polyp Segmentation

Loris Nanni<sup>1</sup>, Carlo Fantozzi<sup>1\*</sup> , Andrea Loreggia<sup>2</sup> and Alessandra Lumini<sup>3</sup>

<sup>1</sup> Department of Information Engineering, University of Padova, Padova, Italy; {loris.nanni,carlo.fantozzi}@unipd.it

<sup>2</sup> Department of Information Engineering, University of Brescia, Brescia, Italy; andrea.loreggia@unibs.it

<sup>3</sup> Department of Computer Science and Engineering, University of Bologna, Bologna, Italy; alessandra.lumini@unibo.it

\* Correspondence: carlo.fantozzi@unipd.it; Tel.: +39-049-827-7947

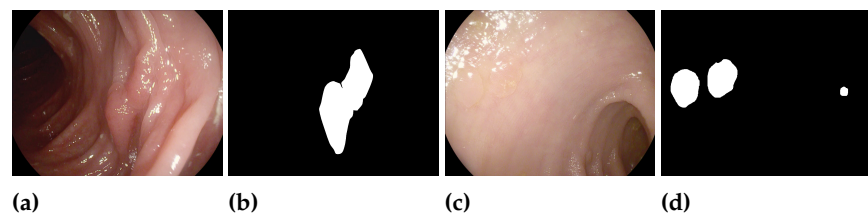
**Abstract:** In the realm of computer vision, semantic segmentation is the task of recognizing objects in images at the pixel level. This is done by performing a classification of each pixel. The task is complex and requires sophisticated skills and knowledge about the context to identify objects' boundaries. 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 we 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 combine 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 which is more suitable for combining transformers in an ensemble. In our extensive experimental evaluation, the proposed ensembles exhibit state-of-the-art performance.

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

## 1. Introduction

Colon polyps are among the preliminary manifestations of colorectal cancer, one of the cancers with the highest incidence [1]. 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]. Nowadays, 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 tireless 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. An example of a colonoscopy image of a polyp and its segmentation is shown in Figure 1, which is taken from [4]. Traditional approaches to segmentation have used, for instance, geometric analysis and a frame-based model [5] or a hybrid context-shape approach [6]. Such approaches hardly extract global context information and are not robust to complex scenes mostly because they rely on hand-crafted features [7]. Deep learning has brought remarkable progress in the field of semantic segmentation and, recently, deep networks have been applied to automatic polyp segmentation in colonoscopy

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.



**Figure 1.** An example of the content of the ETIS-Larib dataset for the semantic segmentation of polyps: (a)(c) original images; (b)(d) ground truth

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 on ensembles for polyp segmentation. As a matter of fact, no earlier review on the topic is available in the open literature, to the best of our knowledge.
- A new ensemble for semantic segmentation based on the HarDNet-MSEG [9], Polyp-PVT [10], and HSNet [11] network topologies. Empirical evaluation shows that performance for polyp segmentation is better than the state of the art.

The idea behind the ensemble is to smooth the contribution of the fine-tuning of the hyper-parameters for a specific dataset while averaging the performance of a model to deal with multiple domains. 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 HSNet, 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 transformers topologies can achieve state-of-the-art performance.

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. Before that, Section 2 contributes a review of the literature on ensembles for polyp segmentation. 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 problem of automatic polyp detection and segmentation. Like in other domains, early attempts (e.g., [7]) focused on detection and are based on features extracted manually. More recently, deep networks have been applied to the task [8]. In recent years, encoder-decoder architectures have gained popularity, with the most widely adopted network in this family being U-Net [12]. Other popular networks adopted for polyp segmentation are reportedly [13] 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 performance. A recent summary of such datasets can be found in [14]. Details about the

datasets we use 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 follow 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 non-medical domains: data augmentation and ensemble techniques.

- Data augmentation [15] 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 [16].
- Ensemble techniques [17] 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 such 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,

- two [18,19] 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 [20–23] have been superseded by newer publications from the same authors, which show better performance on a wider range of datasets.
- Three [24–26] exhibit results that are no longer 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 [27] currently adopted by several researchers, including ourselves.
- Four [13,28–30] report outstanding performance figures, but again, they are obtained with less stringent and/or not thoroughly documented protocols. Except for [30], the source code is not available.

The remaining two works [31,32] are included in the comparisons performed 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 [13], the polyp segmentation method is based on an ensemble built from three different networks: U-Net, SegNet, and PSPNet. The final segmentation is obtained through per-pixel weighted voting of the outputs of the three networks. Weights are proportional to the performance of the networks measured in the validation phase. Training, validation, and testing are performed with images from three public datasets: CVC-ColonDB [7] (300 images), CVC-ClinicDB [4] (612 images), and ETIS-Larib [33] (196 images). The training phase relies on transfer learning and data augmentation (scaling, flipping, rotations at different angles, and changes in brightness).

In [24,34], an ensemble of two Mask R-CNN models with different encoding backbones (ResNet-50 and ResNet-101 [35]) is proposed. The final segmentation is the bitwise combination, using the union operator, of the outputs of the two subnetworks. Transfer learning was adopted: the networks are pre-trained on the COCO dataset and fine-tuned

with images from the same three datasets considered in [13]. Like in [13], data augmentation (scaling, flipping, cropping, padding, random rotations, random shearing, random Gaussian blurring, random contrast normalization, and random changes in brightness) is adopted during training. The authors state that this was an effective tool for improving segmentation performance, confirming the common conclusion in the literature.

In [28], the authors address the segmentation of multiple anatomical structures, including polyps. The proposed segmentation ensemble combines three DeepLabv3+ [36] variants trained with images at different resolutions and with different dilation strides. The authors state that in this way the ensemble captures information at multiple scales. Furthermore, a novel loss function is 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 are 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 is trained and tested with images from the CVC-ColonDB, CVC-ClinicDB, and ETIS-Larib datasets. The datasets are augmented with standard geometric alterations (reflection, random cropping, translation rotation), elastic distortions, contrast normalization, and boundary enhancement.

In [25], the segmentation network combines the predictions of two U-Net models with ResNet-34 [35] and EfficientNet B2 [37] as their backbones; details of how the outputs of the two models are combined to provide the final segmentation are not reported. The networks are trained with transfer learning, with initial weights obtained from ImageNet and fine-tuning performed with the publicly available Kvasir-SEG dataset [38] (1000 images). The segmentation accuracy is then tested with 160 images from the MediEval 2020 challenge. Like in the previous works, the training phase leverages data augmentation (scaling, flipping, random rotations, affine transformations, elastic deformations, CutMix regularization, random changes in contrast, and addition of Gaussian noise). The authors observe 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 [29], different networks (namely, MobileNet, ResNet, and EfficientNet [37]) are first tested as backbones of U-Net, finding that EfficientNet provides the highest performance. Then, a new ensemble is proposed that combines the segmentation results of two U-Nets with EfficientNet B4 and EfficientNet B5 as 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 is adopted during training to take into account the fact that the data are 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  $F_\beta$  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) is employed during training. Transfer learning is also applied: the initial U-Net model is pre-trained on the ImageNet dataset and fine-tuned on the CVC-ClinicDB dataset. The performance of the ensemble is evaluated on the CVC-ColonDB and ETIS-Larib datasets.

Finally, we briefly mention [39] since the authors use the term “ensemble” to describe the networks they examine. However, what they do is test variations of the U-Net architecture with different encoders. Each tested model is a single network, not an ensemble, with a different feature extractor. Based on the results of the experiments, the best-performing feature extractors are DenseNet169 and InceptionResNetV2. Like in several other studies, extensive data augmentation (flipping, blurring, sharpening, random changes in contrast and brightness) is applied during training.

All the aforementioned works combine 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 in a hierarchical way or in some other complex fashion.

In [30], two different ensembles for the semantic segmentation of polyps are discussed. The first ensemble, named TriUNet by the authors, combines three U-Net networks. The second ensemble, called DivergentNets, combines TriUNet with UNet++ [40], FPN [41], DeepLabv3, and DeepLabv3+. The final segmentation mask is an average of the five masks provided by these networks. DivergentNets can be considered an ensemble of size 8, albeit the outputs of three networks are pre-combined in TriUNet. In [22], several ensembles are tested whose components differ in the backbones adopted, the loss functions, and the optimizers used in the training phase. The base architectures for the networks that make up the ensembles are DeepLabv3+ and HarDNet-MSEG [9]. The size of the ensembles ranges from two to 60. The ensembles are 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 [42]. This is an experimental protocol that was first introduced in [27]. In [31], the authors propose ensembles of DeepLabv3+, HarDNet-MSEG, and Polyp-PVT [10] networks, trained with different loss functions and data augmentation methods. A wider range of loss functions is considered than in [22], including weighted combinations of base functions, and more than ten data augmentation techniques are applied in the training phase. The size of the ensembles ranges from two to 14. The datasets used for training and testing are the same as in [27]. In [19], an ensemble is proposed that combines the predictions of Eff-UNet [43], nnU-Net [44], and a hierarchical multi-scale attention network [45]. The training set is 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 has been 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 are on “folds” of data that do not have a documented relationship with public datasets. In [32], the ensemble is 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 is provided by varying the activation functions (15 different loss functions are considered) and the data augmentation strategies. Additionally, in the sub-ensemble based on DeepLabv3+, polyps identified by different networks are not allowed to overlap. The training and testing protocols are, once again, those introduced in [27].

### 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 performance.

#### 3.1. Structure of the Ensemble

As anticipated in Section 1, our ensemble is based on the HarDNet-MSEG [9], Polyp-PVT [10], and HSNet [11] network topologies. In our preliminary experiments, we observed that in these models, the last sigmoid layer followed by a normalization layer has a negative impact on 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 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 than 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



**Table 1.** Ensembles for polyp segmentation: structure of the ensembles. The paper [39] does not appear in this table and in the following. The authors use the term ensemble to describe the networks they examine. However, what they do is test variations of the U-Net architecture with different encoders.

Paper	Ensemble Size	Deep Labv3	Deep Labv3+	Eff-UNet	FPN	HardNet-MSEG	Mask R-CNN	MultiRes UNet	nnU-Net	PSPNet	Polyp-PVT	SegNet	U-Net	Unet++
Guo2019 [13]	3									✓		✓	✓	
Kang2019 [24]	2						✓							
Nguyen2019 [28]	3		✓											
Shrestha2020 [25]	2												✓	
ThuHong2020 [29]	2												✓	
Hong2021 [26]	5													
Lumini2021a [20]	14		✓				✓							
Lumini2021b [21]	14		✓											
Nanni2021 [22]	2, 10, 20, 30, 60		✓			✓								✓
Thambawita2021 [30]	3, 7	✓			✓								✓	
Tomar2021 [18]	4							✓						
Nanni2022a [31]	2, 10, 14		✓			✓					✓			
Nanni2022b [23]	12, 14, 15, 16, 20		✓			✓								
Tran2022 [19]	N/A			✓					✓					
Nanni2023 [32]	2, 4, 10, 20, 30	✓	✓			✓								

**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 [13]	✓	✓				✓			
Kang2019 [24]	✓	✓				✓			
Nguyen2019 [28]	✓	✓				✓			
Shrestha2020 [25]								✓	✓
ThuHong2020 [29]	✓	✓				✓			
Hong2021 [26]					✓				
Lumini2021a [20]								✓	
Lumini2021b [21]								✓	
Nanni2021 [22]	✓	✓	✓			✓		✓	
Thambawita2021 [30]					✓		✓		
Tomar2021 [18]					✓				
Nanni2022a [31]	✓	✓	✓			✓		✓	
Nanni2022b [23]	✓	✓	✓			✓		✓	
Tran2022 [19]	✓	✓		✓	✓	✓			
Nanni2023 [32]	✓	✓	✓			✓		✓	

**Table 3.** Ensembles for polyp segmentation: performance metrics. “FPS” stands for “frames per second”. The authors of [18] provide 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 [13]		✓			✓				
Kang2019 [24]					✓	✓	✓		
Nguyen2019 [28]	✓	✓				✓	✓	✓	✓
Shrestha2020 [25]	✓	✓	✓	✓	✓	✓	✓		
ThuHong2020 [29]		✓			✓	✓	✓		
Hong2021 [26]	✓	✓	✓	✓	✓	✓	✓		
Lumini2021a [20]	✓	✓	✓		✓	✓	✓		
Lumini2021b [21]	✓	✓	✓		✓	✓	✓		
Nanni2021 [22]		✓			✓				
Thambawita2021 [30]		✓			✓	✓	✓		
Tomar2021 [18]									
Nanni2022a [31]		✓			✓				
Nanni2022b [23]	✓	✓	✓		✓	✓	✓		
Tran2022 [19]		✓							
Nanni2023 [32]		✓			✓				

**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 results from different papers are, in general, not comparable because they are 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 [13]			0.967	0.983	0.962	0.980	0.970	0.985						
Kang2019 [24]					0.695		0.661							
Nguyen2019 [28]				0.891		0.896								
Shrestha2020 [25]	0.760	0.838											0.755	0.832
ThuHong2020 [29]					0.798	0.891	0.702	0.823						
Hong2021 [26]											0.571	0.619		
Lumini2021a [20]	0.820	0.885												
Lumini2021b [21]	0.825	0.888												
Nanni2021 [22]	0.872	0.919	0.886	0.931	0.701	0.776	0.663	0.743	0.831	0.901				
Thambawita2021 [30]											0.840			
Nanni2022a [31]	0.874	0.920	0.894	0.937	0.751	0.826	0.717	0.787	0.842	0.904				
Nanni2022b [23]	0.870	0.918	0.884	0.929	0.695	0.768	0.644	0.727	0.833	0.904				
Nanni2023 [32]	0.871	0.920	0.903	0.947	0.720	0.787	0.688	0.756	0.846	0.909				

summed and then passed to the sigmoid, we pass each mask separately to the sigmoid and average 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.

### 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 [46].

We adopt the Dice coefficient [47] 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

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

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

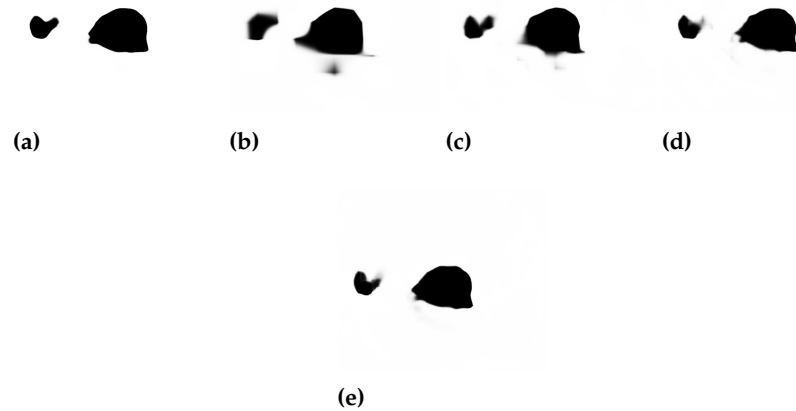
Another well-known performance measure is the intersection over union (IoU), which was introduced for the first time in [48]:

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

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

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





**Figure 2.** Masks in HSNNet. (a): final segmentation mask. (b)-(e) intermediate masks masks after the sigmoid computation.

This could be an issue when dealing with imbalanced data sets. Therefore, as suggested in [49], we use 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 pixel,  $K$  is the number of classes,  $w_{ik}$  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 in order 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 data sets. Hence, the weighted CE loss was introduced and it resulted in well-balanced classifiers for imbalanced scenarios [50]. 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 pixel,  $K$  is the number of classes,  $w_{ik}$  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  $31 \times 31$  and a stride of 1 in order 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 [9], the wIoU loss and the weighted binary CE loss are considered together (structure loss function):

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

In our experiments, the structure loss function is used to train all the networks with the exception of DeepLabv3+.

### 3.3. Datasets

Polyp segmentation from colonoscopy images is a challenging task that requires a 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 [51] 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 720x576 up to 1920x1072 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 [42].
- The ETIS-Larib [33] dataset (“ETIS”) contains 196 colonoscopy images.
- CVC-ClinicDB [4] (“ClinDB”) contains 612 images from 31 different 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 576x768 pixels in size. A frame from the dataset and its corresponding ground truth masks is shown in Figure 1.

Our training set is made by 1450 images taken from the largest datasets, i.e. 900 images from Kvasir and 550 images from ClinDB. The remaining images (100 images from Kvasir, 380 from ColDB, 60 from CVC-T, 196 from ETIS and 62 from ClinDB) are for the test sets, as usually done in the literature. Having a small training dataset 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 adopt two common techniques for addressing this issue: fine-tuning and data augmentation. Fine-tuning is used since all the models involved in our experiments are pre-trained on a large dataset. Data augmentation is based on two different strategies (see Section 3.4), which aim to increase the effective size of the training set and provide the model with more examples to learn from.

### 3.4. Data Augmentations

We consider two data augmentation strategies.

- “DA1”: a basic strategy that includes only image flips and rotations.
- “DA2”: a sophisticated strategy that creates artificial images in 11 different ways, including the application of motion blur and shadows to original images.

The two strategies were introduced in [31]: we refer the interested reader to this work for a thorough description.

### 3.5. Overview of the Experiments

In our experiments we consider 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 [52] in the domain of polyp segmentation. The networks are 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 have used the default value for the number of epochs. The networks are always trained using the structure loss function (see Section 3.2) except for DeepLabv3+, and their outputs are combined with the sum rule or the weighted sum rule depending on the experiment. We use resized training images of size 352x352. During the test phase, masks are obtained by resizing the images, and they are subsequently resized back to the original dimensions to evaluate the performance of the model.

**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	<b>0.918</b>	<b>0.947</b>	0.778	0.756	<b>0.909</b>	0.862
		b	0.907	0.928	0.775	0.770	0.899	0.856
		a+b	0.915	0.931	<b>0.785</b>	<b>0.781</b>	0.904	<b>0.863</b>

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

We train the networks with the two data augmentation techniques described in Section 3.4. We also experiment with training two identical networks with the two techniques, and combine their outputs.

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

- $10^{-4}$  (learning rate “a”),
- $5 \cdot 10^{-5}$  decaying to  $5 \cdot 10^{-6}$  after 30 epochs (learning rate “b”).

We also train some networks twice with the two learning rates and, like we do with the data augmentations, combine their outputs.

#### 4. Experimental Results

In this section, we report on the experimental analysis carried out to assess the proposed ensemble strategy. In the first set of experiments, we separately analyze the potential of each of the three network topologies, that is, HarDNet-MSEG [9], Polyp-PVT [10], and HSNet [11]. The results are summarized in Tables 5, 6, and 7, respectively. The Dice coefficient is used to measure performance on the five test sets defined in Section 3.3. In column DA we specify which data augmentation approaches (see Section 3.4) are considered. Where the column states “1+2” we trained the network twice with data augmentations 1 and 2. In column LR we specify which learning rates (see Section 3.5) are 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 3.5) are 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 is, therefore, an ensemble. Consequently, the last three columns in the table represent ensembles of 4, 4, and 8 HarDNet-MSEG networks, respectively.

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

**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	ColDB	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	<b>0.931</b>	0.920	0.792	0.776	0.876	0.859
1+2	b	Yes	<b>0.931</b>	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	<b>0.933</b>	<b>0.824</b>	<b>0.808</b>	<b>0.895</b>	<b>0.877</b>

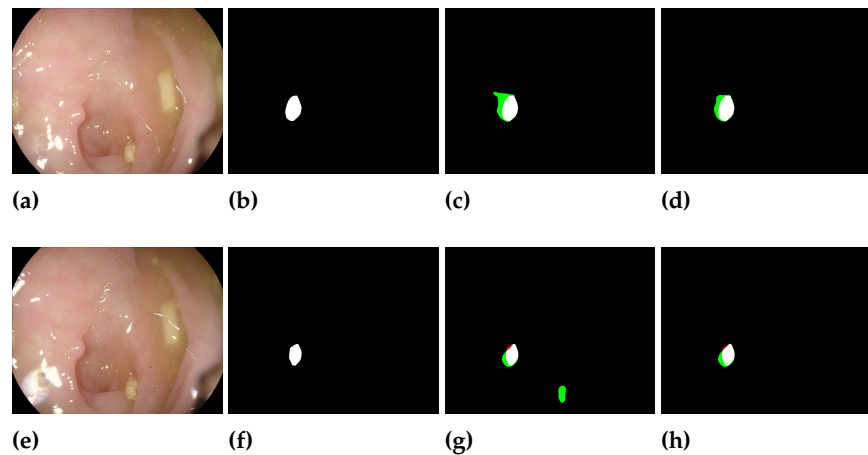
**Table 7.** Dice coefficient for the (ensemble) networks built on HSNet. The best performance figures are marked in bold.

DA	LR	SM	Kvasir	ClinDB	ColDB	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	<b>0.947</b>	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	<b>0.931</b>	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	<b>0.829</b>	0.779	0.902	0.876
1+2	a+b	Yes	0.928	0.945	0.828	<b>0.791</b>	<b>0.905</b>	<b>0.879</b>

In Tables 6 and 7, the additional column SM specifies whether we obtain the final segmentation mask like in the original HSNet and Polyp-PVT networks (SM = “No”) or with the novel approach we propose in Section 3.1 (SM = “Yes”). The conclusions we can draw from the results reported in Tables 6 and 7 are that, on average, learning rate b performs 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 is obtained by the ensemble obtained by varying the data augmentation and the learning rate.

Using the proposed approach to obtain the segmentation masks allows 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 3, where false-positive pixels are highlighted in green, while the false-negatives are in red. They demonstrate that our ensemble model with smoothing produces better boundary results and makes more accurate predictions with respect to one 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. It can be seen that these output values are already very close to saturating the sigmoid function, thus summing them together would make the situation worse by producing an almost binary output. Our approach, instead, averages the intermediate masks and produces a smoother output, which means that it maintains more information. This information can be visually appreciated in Figure 3. In particular, it can be noticed that the final segmentation mask of HSNet (Figure 3 (a)) is very sharp, indicating an almost binary output, while the four intermediate masks  $P_1$ - $P_4$  have more blurred edges.



**Figure 3.** Masks in HSNet ensemble with and without smoothing. (a)(e): original image, (b)(f) ground truth, (c)(g): HSNet ensemble without smoothing, (d)(h): HSNet ensemble with smoothing. False-positive pixels are in green, while the false-negatives are in red.

In Table 9, our ensembles are compared with several approaches reported in the literature. In our final proposed ensembles, the methods are combined not with the sum rule, but with the weighted sum rule. Each method is 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 8 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 “ $FH(2) + 2 \times PVT(2)$ ”, the best ensemble introduced in [31] and based on DeepLabv3+.

From the results reported in Table 9, it can be noticed that the proposed ensembles beat the state of the art on the ColDB and ETIS datasets. We remark that this result is obtained without training the ensembles specifically on such datasets. Most importantly, all three proposed ensembles perform better than the state of the art when we average across all datasets. 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 are strong performers with all the datasets: even when they are not the best, they are near the top. This is a benefit of the ensemble strategy. Of course, the three ensembles perform significantly better than their composing networks (Table 9, rows 4 to 7), including the recent ensemble  $FH(2) + 2 \times PVT(2)$ .

## 5. Conclusion

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 in order to improve the overall performance. We introduced an effective way of doing this by averaging the predictions of the individual models 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.

**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.

Method	Min_max	Avg_max	Avg_min	Max_min
HarDNet-MSEG	2.51	71.84	-6.53	-11.45
Polyp-PVT: $P_1$	26.22	70.15	-21.06	-30.04
Polyp-PVT: $P_2$	19.53	38.55	-20.61	-29.47
HSNet: $P_1$	24.21	67.81	-38.49	-69.06
HSNet: $P_2$	33.19	87.04	-42.63	-84.26
HSNet: $P_3$	31.31	122.01	-63.13	-123.45
HSNet: $P_4$	36.47	160.11	-95.23	-165.40

**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 are 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
<i>Ens1</i>	<b>0.886</b>	0.930	0.892	0.936	<b>0.764</b>	<b>0.839</b>	<u>0.738</u>	<u>0.812</u>	0.841	0.904	<u>0.824</u>	<u>0.884</u>
<i>Ens2</i>	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>	<b>0.885</b>
<i>Ens3</i>	0.882	0.928	0.894	0.937	<u>0.752</u>	0.824	<b>0.760</b>	<b>0.830</b>	0.840	0.905	<b>0.826</b>	<b>0.885</b>
HarDNet-MSEG [9]	0.857	0.912	0.882	0.932	0.66	0.731	0.613	0.677	0.821	0.887	0.767	0.828
Polyp-PVT [10]	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 [11]	0.877	0.926	<b>0.905</b>	<b>0.948</b>	0.735	0.81	0.734	0.808	0.839	0.903	0.818	0.879
$FH(2) + 2 \times PVT(2)$ [31]	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 [9])	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 [9])	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 [9])	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 [9])	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 [32]	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 [53]	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 [54]	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 [55]	<b>0.886</b>	<b>0.932</b>	0.886	0.933	0.73	0.803	0.711	0.79	<b>0.859</b>	<b>0.919</b>	0.814	0.875
SETR [56]	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 [57]	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 [58]	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 [59]	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 [60]	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 [61]	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 [62]	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 [63]	0.865	0.912	0.888	0.936	0.69	0.762	0.679	0.756	-	-	-	-



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 learning rate strategy is a feasible method to build a set of segmentation networks;
- a better way to add the transformers (Polyp-PVT and HSNet) 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 Al.Lu.; writing—original draft preparation, C.F., An.Lo. and L.N.; writing—review and editing, C.F., An.Lo., L.N. and Al.Lu. 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>.

**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: A Cancer Journal for Clinicians* **2020**, *70*, 145–164. <https://doi.org/10.3322/CAAC.21601>.
2. Valderrama-Treviño, A.; Hazzel, E.; Flores, C.; Herrera, M. Colorectal cancer: a review. *Article in International Journal of Research in Medical Sciences* **2017**. <https://doi.org/10.18203/2320-6012.ijrms20174914>.
3. Wieszczy, P.; Regula, J.; Kaminski, M.F. Adenoma detection rate and risk of colorectal cancer. *Best Practice & Research Clinical Gastroenterology* **2017**, *31*, 441–446. <https://doi.org/10.1016/j.BPG.2017.07.002>.
4. 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. *Computerized Medical Imaging and Graphics* **2015**, *43*, 99–111. <https://doi.org/10.1016/j.compmedimag.2015.02.007>.
5. Mamonov, A.V.; Figueiredo, I.N.; Figueiredo, P.N.; Richard Tsai, Y.H. Automated polyp detection in colon capsule endoscopy. *IEEE Transactions on Medical Imaging* **2014**, *33*, 1488–1502, [1305.1912]. <https://doi.org/10.1109/TMI.2014.2314959>.
6. Tajbakhsh, N.; Gurudu, S.R.; Liang, J. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE Transactions on Medical Imaging* **2016**, *35*, 630–644. <https://doi.org/10.1109/TMI.2015.2487997>.
7. Bernal, J.; Sánchez, J.; Vilariño, F. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* **2012**, *45*, 3166–3182. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), <https://doi.org/10.1016/j.patcog.2012.03.002>.
8. Pacal, I.; Karaboga, D.; Basturk, A.; Akay, B.; Nalbantoglu, U. A comprehensive review of deep learning in colon cancer. *Computers in Biology and Medicine* **2020**, *126*, 104003. <https://doi.org/10.1016/j.combiomed.2020.104003>.
9. 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, 2021. <https://doi.org/10.48550/ARXIV.2101.07172>.
10. Dong, B.; Wang, W.; Fan, D.P.; Li, J.; Fu, H.; Shao, L. Polyp-PVT: Polyp Segmentation with Pyramid Vision Transformers, 2021. <https://doi.org/10.48550/ARXIV.2108.06932>.
11. Zhang, W.; Fu, C.; Zheng, Y.; Zhang, F.; Zhao, Y.; Sham, C.W. HSNet: A hybrid semantic network for polyp segmentation. *Computers in Biology and Medicine* **2022**, *150*, 106173. <https://doi.org/10.1016/j.combiomed.2022.106173>.
12. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; Navab, N.; Hornegger, J.; Wells, W.M.; Frangi, A.F., Eds.; Springer International Publishing: Cham, 2015; pp. 234–241.
13. 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. *Medical Physics* **2019**, *46*, 5666–5676. <https://doi.org/10.1002/mp.13865>.

14. 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. *Machine Intelligence Research* **2022**, pp. 1–19. <https://doi.org/10.1007/s11633-022-1371-y>.
15. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *Journal of big data* **2019**, *6*, 1–48. <https://doi.org/10.1186/s40537-019-0197-0>.
16. Singh, N.K.; Raza, K. Medical image generation using generative adversarial networks: A review. *Health informatics: A computational perspective in healthcare* **2021**, pp. 77–96.
17. Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A survey on ensemble learning. *Frontiers of Computer Science* **2020**, *14*, 241–258. <https://doi.org/10.1007/s11704-019-8208-z>.
18. 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 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), 2021, Vol. 2886, pp. 49–58.
19. 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 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), 2022, Vol. 3148, pp. 20–24.
20. Lumini, A.; Nanni, L.; Maguolo, G. Deep ensembles based on Stochastic Activation Selection for Polyp Segmentation. In Proceedings of the Medical Imaging with Deep Learning, 2021.
21. Lumini, A.; Nanni, L.; Maguolo, G. Deep Ensembles Based on Stochastic Activations for Semantic Segmentation. *Signals* **2021**, *2*, 820–833. <https://doi.org/10.3390/signals2040047>.
22. Nanni, L.; Cuza, D.; Lumini, A.; Loreggia, A.; Brahnam, S. Deep ensembles in bioimage segmentation, 2021. <https://doi.org/10.48550/ARXIV.2112.12955>.
23. Nanni, L.; Cuza, D.; Lumini, A.; Brahnam, S. Data augmentation for deep ensembles in polyp segmentation. In *Computational Intelligence Based Solutions for Vision Systems*; 2053-2563, IOP Publishing, 2022; pp. 8–1 to 8–22. <https://doi.org/10.1088/978-0-7503-4821-8ch8>.
24. Kang, J.; Gwak, J. Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images. *IEEE Access* **2019**, *7*, 26440–26447. <https://doi.org/10.1109/ACCESS.2019.2900672>.
25. Shrestha, S.; Khanal, B.; Ali, S.; et al. Ensemble U-Net Model for Efficient Polyp Segmentation. In Proceedings of the Proceedings of the MediaEval 2020 Workshop, 2020, Vol. 2882.
26. Hong, A.; Lee, G.; Lee, H.; Seo, J.; Yeo, D. Deep Learning Model Generalization with Ensemble in Endoscopic Images. In Proceedings of the 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), 2021, Vol. 2886, pp. 80–89.
27. 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 Proceedings of the 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 International Publishing: Cham, 2020; pp. 263–273. [https://doi.org/10.1007/978-3-030-59725-2\\_26](https://doi.org/10.1007/978-3-030-59725-2_26).
28. 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. <https://doi.org/10.1109/ACCESS.2019.2904094>.
29. 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), 2020, pp. 1–6. <https://doi.org/10.1109/RIVF48685.2020.9140793>.
30. Thambawita, V.; Hicks, S.; Halvorsen, P.; Riegler, M. DivergentNets: Medical Image Segmentation by Network Ensemble. In Proceedings of the 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), 2021, Vol. 2886, pp. 27–38.
31. Nanni, L.; Lumini, A.; Loreggia, A.; Formaggio, A.; Cuza, D. An Empirical Study on Ensemble of Segmentation Approaches. *Signals* **2022**, *3*, 341–358. <https://doi.org/10.3390/signals3020022>.
32. 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, 2023; pp. 133–153. [https://doi.org/10.1007/978-3-031-11154-9\\_7](https://doi.org/10.1007/978-3-031-11154-9_7).
33. 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**. <https://doi.org/10.1007/s11548-013-0926-3>.
34. Kang, J.; Gwak, J. Corrections to “Ensemble of Instance Segmentation Models for Polyp Segmentation in Colonoscopy Images”. *IEEE Access* **2020**, *8*, 100010–100012. <https://doi.org/10.1109/ACCESS.2020.2995611>.
35. 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), 2016, pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
36. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the Computer Vision – ECCV 2018; Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y., Eds.; Springer International Publishing: Cham, 2018; pp. 833–851. [https://doi.org/10.1007/978-3-030-01234-2\\_49](https://doi.org/10.1007/978-3-030-01234-2_49).

37. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning; Chaudhuri, K.; Salakhutdinov, R., Eds. PMLR, 2019, Vol. 97, *Proceedings of Machine Learning Research*, pp. 6105–6114.
38. 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 Proceedings of the 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 International Publishing: Cham, 2020; pp. 451–462. [https://doi.org/10.1007/978-3-030-37734-2\\_37](https://doi.org/10.1007/978-3-030-37734-2_37).
39. Hosseinzadeh Kassani, S.; Hosseinzadeh Kassani, P.; Wesolowski, M.J.; Schneider, K.A.; Deters, R. Automatic Polyp Segmentation Using Convolutional Neural Networks. In Proceedings of the Advances in Artificial Intelligence; Goutte, C.; Zhu, X., Eds.; Springer International Publishing: Cham, 2020; pp. 290–301. [https://doi.org/10.1007/978-3-030-47358-7\\_29](https://doi.org/10.1007/978-3-030-47358-7_29).
40. Zhou, Z.; Rahman Siddiquee, M.M.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Proceedings of the 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 International Publishing: Cham, 2018; pp. 3–11. [https://doi.org/10.1007/978-3-030-00889-5\\_1](https://doi.org/10.1007/978-3-030-00889-5_1).
41. 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), 2017, pp. 936–944. <https://doi.org/10.1109/CVPR.2017.106>.
42. 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. *Journal of healthcare engineering* **2017**, 2017. <https://doi.org/10.1155/2017/4037190>.
43. 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), 2020, pp. 1473–1481. <https://doi.org/10.1109/CVPRW50498.2020.00187>.
44. 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. *Nature methods* **2021**, 18, 203–211. <https://doi.org/10.1038/s41592-020-01008-z>.
45. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation, 2020. <https://doi.org/10.48550/ARXIV.2005.10821>.
46. Jadon, S. A survey of loss functions for semantic segmentation.
47. 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.
48. Rahman, M.A.; Wang, Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In Proceedings of the International symposium on visual computing. Springer, 2016, pp. 234–244.
49. Cho, Y.J. Weighted Intersection over Union (wIoU): A New Evaluation Metric for Image Segmentation. *arXiv preprint arXiv:2107.09858* **2021**.
50. Aurelio, Y.S.; de Almeida, G.M.; de Castro, C.L.; Braga, A.P. Learning from imbalanced data sets with weighted cross-entropy function. *Neural processing letters* **2019**, 50, 1937–1949.
51. 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 Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, pp. 164–169.
52. Shen, T.; Xu, H. Medical image segmentation based on Transformer and HardNet structures. *IEEE Access* **2023**. <https://doi.org/10.1109/ACCESS.2023.3244197>.
53. Li, W.; Zhao, Y.; Li, F.; Wang, L. MIA-Net: Multi-information aggregation network combining transformers and convolutional feature learning for polyp segmentation. *Knowledge-Based Systems* **2022**, 247, 108824. <https://doi.org/10.1016/j.knosys.2022.108824>.
54. Wu, Y.H.; Liu, Y.; Zhan, X.; Cheng, M.M. P2T: Pyramid Pooling Transformer for Scene Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**, pp. 1–12. <https://doi.org/10.1109/tpami.2022.3202765>.
55. Liu, F.; Hua, Z.; Li, J.; Fan, L. DBMF: Dual Branch Multiscale Feature Fusion Network for polyp segmentation. *Computers in Biology and Medicine* **2022**, 151, 106304. <https://doi.org/10.1016/j.compbiomed.2022.106304>.
56. 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), 2021, pp. 6877–6886. <https://doi.org/10.1109/CVPR46437.2021.00681>.
57. 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, 2021. <https://doi.org/10.48550/ARXIV.2102.04306>.
58. Zhang, Y.; Liu, H.; Hu, Q. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. In Proceedings of the 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 International Publishing: Cham, 2021; pp. 14–24. [https://doi.org/10.1007/978-3-030-87193-2\\_2](https://doi.org/10.1007/978-3-030-87193-2_2).

- 
59. Kim, T.; Lee, H.; Kim, D. UACANet: Uncertainty Augmented Context Attention for Polyp Segmentation. In Proceedings of the Proceedings of the 29th ACM International Conference on Multimedia; Association for Computing Machinery: New York, NY, USA, 2021; MM '21, pp. 2167–2175. <https://doi.org/10.1145/3474085.3475375>.
  60. Wei, J.; Hu, Y.; Zhang, R.; Li, Z.; Zhou, S.K.; Cui, S. Shallow Attention Network for Polyp Segmentation. In Proceedings of the Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021, Vol. 12901 LNCS. [https://doi.org/10.1007/978-3-030-87193-2\\_66](https://doi.org/10.1007/978-3-030-87193-2_66).
  61. Zhao, X.; Zhang, L.; Lu, H. Automatic Polyp Segmentation via Multi-scale Subtraction Network, 2021. <https://doi.org/10.48550/ARXIV.2108.05082>.
  62. Park, K.B.; Lee, J.Y. SwinE-Net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and Swin Transformer. *Journal of Computational Design and Engineering* **2022**, *9*, 616–632. <https://doi.org/10.1093/jcde/qwac018>.
  63. Song, P.; Li, J.; Fan, H. Attention based multi-scale parallel network for polyp segmentation. *Computers in Biology and Medicine* **2022**, *146*, 105476. <https://doi.org/10.1016/j.combiomed.2022.105476>.

**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.