Towards Automated Polyp Segmentation Using Weakly- and Semi-Supervised Learning and Deformable Transformers

Guangyu Ren* Michalis Lazarou* Jing Yuan* Tania Stathaki
Imperial College London

Abstract

Polyp segmentation is a crucial step towards computer-aided diagnosis of colorectal cancer. However, most of the polyp segmentation methods require pixel-wise annotated datasets. Annotated datasets are tedious and time-consuming to produce, especially for physicians who must dedicate their time to their patients. To this end, we propose a novel weakly- and semi-supervised learning polyp segmentation framework that can be trained using only weakly annotated images along with unlabeled images making it very cost-efficient to use. More specifically our contributions are: 1) a novel weakly annotated polyp dataset, 2) a novel sparse foreground loss that suppresses false positives and improves weakly-supervised training, 3) a deformable transformer encoder neck for feature enhancement by fusing information across levels and flexible spatial locations.

Extensive experimental results demonstrate the merits of our ideas on five challenging datasets outperforming some state-of-the-art fully supervised models. Also, our framework can be utilized to fine-tune models trained on natural image segmentation datasets drastically improving their performance for polyp segmentation and impressively demonstrating superior performance to fully supervised fine-tuning. Code can be found in https://github.com/ic-qialanqian/WS-DefSegNet.

1. Introduction

Automated medical image segmentation has attracted interest in recent years due to its potential to significantly reduce the workload of physicians by being used as a supporting tool for a physician's diagnosis. Due to the rapid development of deep learning [14], the current state-of-the-art medical image segmentation methods utilize deep learning techniques and polyp segmentation has been no exception [11, 12, 37].

However, one of the bottlenecks of deep learning techniques is their reliance on large, well-annotated datasets. Annotating datasets for polyp segmentation is particularly time-consuming since pixel-wise annotations must be provided which requires significant manual labour. While in standard image segmentation anyone can annotate a dataset, in polyp images, annotations must be provided by expert physicians that are trained to detect lesions in these images. This is a significant limitation for automated polyp segmentation since physicians do not have time to dedicate to annotating images.

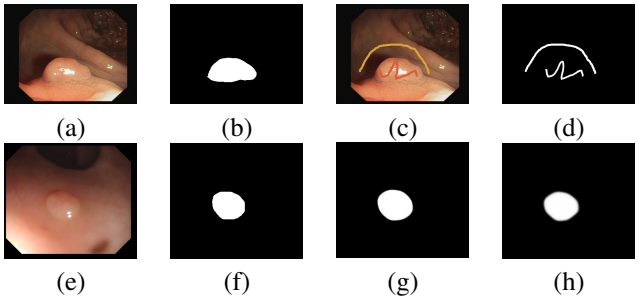


Figure 1. Visualization of weak annotations. (a) RGB image of the training data. (b) Original ground truth. (c) Foreground and background. (d) Our weak annotations. (e) RGB image of the testing data. (f) Corresponding ground truth. (g) Predicted map of fully supervised training manner. (h) Predicted map of our methods.

To address this issue and save physicians' valuable time we propose a novel framework for polyp segmentation. Our framework can be trained using only weakly annotated images and unlabeled images. These weak annotations include only information regarding where the foreground and background pixels are located.

Specifically, we leverage our framework on polyp segmentation, which aims at detecting and segmenting polyps for the early diagnosis of colorectal cancer. Current research [11, 12, 37] still relies on complete polyp annotations to achieve accurate detection performance. Under this circumstance, we relabel the training dataset with weak annotations by simply drawing sketches. Only around 1.9% of the total pixels of all images in the whole dataset are labeled. The annotations simply need to indicate the foreground (polyp region) and the background (non-polyp region), making this annotation strategy very efficient for physicians to use without sacrificing a lot of time. Our weak annotations can be

seen in Figure 1(c) and (d) annotating the polyp region and the non-polyp region with two simple lines in direct contrast to the original ground truth segmentation maps that require pixel-wise careful annotation.

Our proposed framework consists of a two-stage training regime. In the first stage, a model is trained using a weakly-supervised training paradigm while in the second stage we train the model using a semi-supervised learning paradigm. Also, as part of our framework, we propose a novel architectural component that is used for feature enhancement.

During the weakly-supervised training stage, we propose a novel weakly-supervised loss function that addresses a key limitation of weakly-supervised training techniques, that of numerous false positives [40, 42]. Since weak annotations contain only a fraction of the polyp region, training models by partial cross-entropy loss [32] could cause a large number of false positives as shown in Figure 2(c). A previous work [42] attempted to address the problem of false positives using an auxiliary edge detection network supervising the model to align image edges with the predicted segmentation map boundaries. However, this method complicates the training process and relies on auxiliary networks. To address this problem in a simple way, we propose a novel sparse foreground loss function that suppresses false positives and refines the rough predicted segmentation maps (Figure 2(d)).

In addition, because of the weakly-supervised training, inconsistent segmentation maps can be generated by two identical models trained the same way (Figure 8(b) and (c)). To exploit the prior knowledge of the predicted map, we adopt a batch-wise weighted consistency loss to utilize two predicted segmentation maps during semi-supervised training.

Lastly, to improve the accuracy performance even further, we propose a Deformable Transformer Encoder Neck (DTEN), which leverages a multi-scale deformable self-attention encoder along with a novel progressive compensation sequence for feature enhancement. This idea is motivated by [39], which models the topological structure of patterns with graphs and fuses features at calculated positions into new features in the Contextual Pattern Propagation (CPP) module. It is shown that the enhanced features benefit weakly-supervision. Different from CPP, we take advantage of deformable vision transformers for feature enhancement. Such transformers learn the feature positions and weights automatically, making the fusion operation more adaptable to varying conditions like shapes.

The merit of each of our ideas can be visualized from Figure 2 (c)-(f). Each idea consistently improves performance. We name our novel framework Weakly- and Semi-supervised Deformable Segmentation network, in short, **WS-DefSegNet**. To summarize, the contributions of our work are the following:

- We are the first, to the best of our knowledge, to propose a weakly- and semi-supervised training framework for efficient polyp segmentation.
- We propose a novel sparse foreground loss function that alleviates the false positives associated with weakly supervised training.
- We propose DTEN, a novel progressive multi-scale architecture with a self-attention mechanism for feature enhancement that significantly improves the performance of WS-DefSegNet.
- We created the first, to the best of our knowledge, weakly annotated polyp segmentation dataset **W-Polyp**. We are planning on making it publicly available as a way to promote research in this direction.

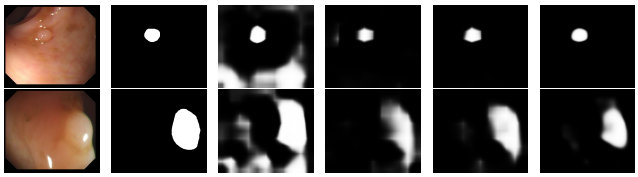


Figure 2. Visual comparison of ablation study. (a) RGB image. (b) Original ground truth (c) backbone. (d) +sparse foreground loss. (e) +semi with weighted consistency loss. (f) +DTEN.

2. Related Work

Medical Image Segmentation Medical image segmentation aims at identifying lesion areas which indicate potential diseases in the human tissue. Deep learning methods have achieved compelling performance due to a fully supervised training paradigm. U-net [28] designs a U-shape architecture built on fully convolutional networks to capture context features and gradually segment biomedical images with precise localization. Analogously, CE-net [15] proposes an encoder-decoder structure with a dense atrous convolution block for medical segmentation and [21] inherits the U-net framework and proposes a non-local context-guided mechanism to capture long-range pixel-wise dependencies in features for tumor segmentation.

More specifically for polyp segmentation, Pranet [11] proposes a recurrent reverse attention module to mine boundary cues and a parallel partial decoder. Other approaches [12, 18, 43] have also been proposed with the overwhelming majority focusing on fully supervised training. In contrast, our framework only uses weak annotations and outperforms some of the aforementioned methods.

Weakly-supervised Segmentation To avoid tediously labeling pixel-wise annotations, image segmentation is en-

couraged through the use of inexpensive labels, formulating the weakly-supervised training paradigm using image-level labels and weak labels. Ahn [2] proposes an IRNet to estimate rough areas of individual instances and detect boundaries with image-level class labels. Chen [9] explicitly explores object boundaries through coarse localization and proposes a BENet to further excavate more object boundaries. Zhang [42] leverages scribble annotations by relabeling an existing salient object detection dataset and further adopting an auxiliary edge detection task to explicitly provide edge supervision on the final output. Yu [40] designs a local coherence loss to improve boundary localization and a structure consistency loss to further enhance the model's generalization ability. However, the aforementioned methods use auxiliary networks and focus on excavating edge information, while in our work we propose an effective weakly-supervised loss function for polyp segmentation.

Semi-supervised Learning Semi-supervised learning addresses the research question of exploiting unlabeled data together with labeled data to improve the performance of a model. A line of research attracting attention in recent years is that of consistency regularization, where the main idea is to enforce similar predictions between two cases, either two different augmentations of the same image or the same image but predictions made from two different networks [19,29,34]. Pseudo-labeling unlabeled data and using them in the training process is another promising direction, for example, [20] uses the current network to assign pseudo-labels to the unlabeled data while [16] uses label propagation to exploit the underlying manifold structure of the data to assign pseudo-labels. Other influential works such as MixMatch [5] and ReMixMatch [4] incorporate many ideas together, such as using data augmentation consistency, applying mixUp regularization [41] and distribution alignment [6]. For further information regarding semi-supervised learning, we refer the reader to [8].

Regarding polyp segmentation, Wu [37] employs two collaborative segmentation networks for semi-supervised polyp segmentation and two discriminators to minimize the impact of the imbalance problem between labeled and unlabeled data. However, in contrast to our work they use a fully annotated subset of polyps while we only use weak annotations.

Vision Transformers in Medical Image Segmentation Vision transformers have been extensively applied to medical image segmentation owing to their capability to incorporate global features while maintaining high resolution. They can be used to establish effective backbones to improve lesion segmentation. [25] stacks four Patcher blocks with vision transformer blocks as the core. [22] encodes input image patches with multiple Swin transformer encoders [24] in parallel with the traditional CNN-based backbone. Besides, transformers are used for feature fusion out of the backbone. [38] combines multi-modality features with the assistance of multiple transformer encoders and a single decoder for MRI brain image segmentation. [36] fuses the patch- and image-level features with three transformer encoders for retinal vessel segmentation. [33] appends six transformer encoder-decoders after the CNN backbone for lesion segmentation. To achieve efficient and accurate segmentation, we take the advantage of the multi-scale deformable transformer [44] and only use a single transformer encoder to maximize inference speed.

3. Efficient Polyp Segmentation

3.1. W-Polyp Dataset

As stated in section 1, we create the first weakly annotated dataset for polyp segmentation comprising of weakly annotated and unlabeled images named W-Polyp. W-Polyp is created by labeling the existing training data of [11] which contains 1,450 images. We randomly selected and weakly annotated 750 images with simple sketches, including lines, scribbles and circles. Annotating an image in this way only takes 2 seconds. Additionally, unlike other weakly annotated datasets, the other 700 images are left unlabeled, maximizing labeling efficiency and enlarging the sparsity of the whole training data. Therefore, only around 1.9% of pixels are labeled as foreground and background as shown in Figure 3.

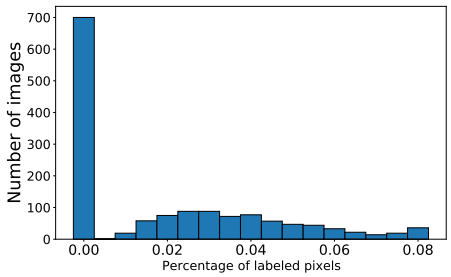


Figure 3. The histogram of the number of images versus the percentage of labeled pixels.

3.2. Method

3.2.1 Overview of WS-DefSegNet

We propose the complete framework named WS-DefSegNet for efficient polyp segmentation. Our framework consists of two-training stages (Figure 4) and a network architecture (Figure 5). The first training stage consists of a weakly-supervised training regime leveraging weakly annotated images while the second stage consists of a semi-supervised training regime leveraging both weakly annotated and unlabeled images. Regarding our network architecture, we

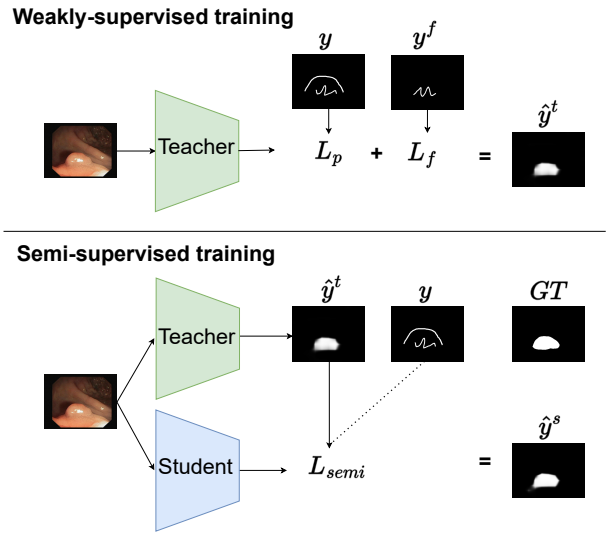


Figure 4. The training procedure of our WS-DefSegNet. We first train the teacher network using a weakly-supervised paradigm as explained in 3.2.3. We train the final student network using a semi-supervised paradigm as explained in 3.2.4. GT denotes the original ground truth map. Our WS-DefSegNet generates a satisfying segmentation map compared to the GT.

propose a novel module, DETN, that uses deformable transformers for feature enhancement.

3.2.2 Problem Formulation

We define the set of all images in our W-Polyp dataset as X. The subset of weakly-annotated images is defined as X_l with their corresponding ground truth maps as Y_l and the subset of unlabeled data as X_u, where $X_l \in X$, $X_u \in X$ and $X_l \cap X_u = \emptyset$. For every batch B, the labeled ground truth including foreground and background information is defined as B_l, while B_l^f denotes only foreground annotations. We denote our model M_θ where θ is the set of learnable parameters. $\hat{y}_i$ is the predicted segmentation map of the i-th image $x_i \in X$, $\hat{y}_i := M_\theta(x_i)$. During the semi-supervised training stage, we use a teacher and a student model. We denote the teacher model as M_θ^t and the student model as M_θ^s.

3.2.3 Weakly-supervised Training

We define the partial cross-entropy loss utilized in [42] as follows:

$$L_p(\hat{y}_i, y_i) = \frac{1}{|B|} \sum_{y_i \in B_l} (y_i log\hat{y}_i + (1-y_i) log(1-\hat{y}_i)) \quad (1)$$

where y_i denotes the corresponding ground truth map with weak sketch annotations. Note that for all of the loss functions used in this work we average all per-pixel losses image-wise but omit this information from the equations to simplify our notation.

In order to mitigate the issue of false positives as described in section 1, we propose a novel loss function that utilizes only the foreground pixels to supervise the model defined as:

$$L_f(\hat{y}_i, y_i^f) = \frac{1}{|B|} \sum_{y_i^f \in B_l^f} (y_i^f log\hat{y}_i + (1-y_i^f) log(1-\hat{y}_i)) \quad (2)$$

where y_i^f indicates a ground truth map with only foreground annotations. Then the total loss for weakly-supervised learning can be defined as:

$$L_{weak}(\hat{y}_i, y_i, y_i^f) = L_p(\hat{y}_i, y_i) + \alpha \cdot L_f(\hat{y}_i, y_i^f) \quad (3)$$

where α is the weight of the sparse foreground loss. It is worth noting that α should be set appropriately. This is because small α makes the predicted segmentation map $\hat{y}_i$ contain many false positives, while large α forces the model to focus on the extremely sparse foreground pixels, leading to more false negatives. In this paper, it is set to 0.5, for further information please refer to the supplementary material.

3.2.4 Semi-supervised Training

Following [34], we adopt a teacher-student learning paradigm and train the teacher model as described in 3.2.3. Using the teacher model, M_θ^t we assign pseudo-labels for every $x_i \in X$ defined as:

$$\hat{y}_i^t = M_\theta^t(x_i) \quad (4)$$

In order to utilize the prior knowledge of the teacher model, M_θ^t, for training the student model, M_θ^s, we utilize a consistency loss for semi-supervised learning:

$$L_c(\hat{y}_i^s, \hat{y}_i^t) = \frac{1}{|B|} \sum_{i \in B} |\hat{y}_i^s - \hat{y}_i^t| \quad (5)$$

where $\hat{y}_i^s$ refers to the predicted map of the student model such that $\hat{y}_i^s := M_\theta^s(x_i)$. For weakly labeled data, the model mainly depends on weakly-supervised learning, and the pseudo labels $\hat{y}_i^t$ can be treated as a regularization term in semi-supervised training. In other words, for every batch B, if there are labeled data in B, namely $X_l \in B$, the training loss is dominated by the L_{weak}. Otherwise, the training loss only depends on the weighted consistency loss L_c. The total loss for semi-supervised training is defined as follows:

$$L_{semi}(\hat{y}_i^s, y_i, y_i^f) = \begin{cases} L_{weak}(\hat{y}_i^s, y_i, y_i^f) + \beta_1 \cdot L_c(\hat{y}_i^s, \hat{y}_i^t) \\ \beta_2 \cdot L_c(\hat{y}_i^s, \hat{y}_i^t) \end{cases} \quad (6)$$

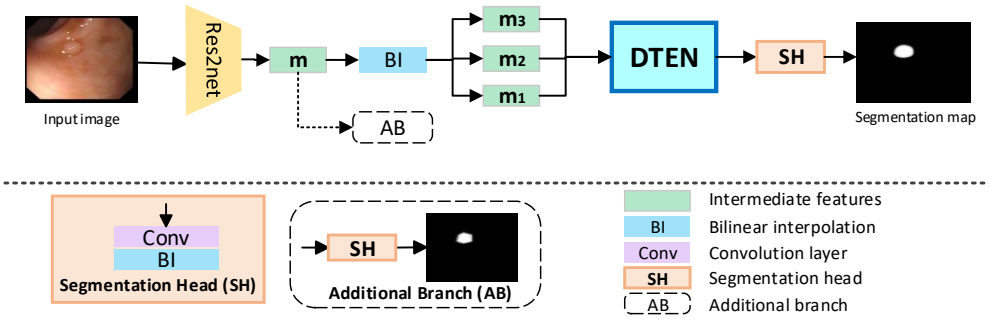


Figure 5. The network architecture of our WS-DefSegNet. It utilizes the proposed Deformable Transformer Encoder Neck (DTEN) to enhance raw features produced by the last stage of the Res2net [13]. Enhanced features are passed to a vanilla segmentation head. The additional branch in the dashed box only exists in the training stage.

Notably, the hyper-parameters β_1 and β_2 in equation 6 are set to 0.1 and 0.5 respectively in this paper. For further information regarding β_1 and β_2 please refer to the supplementary material. Thus, the model is able to refine the final predicted map by considering the prior knowledge of the first rough predicted map. The overall training procedure is illustrated in Figure 4.

3.2.5 Deformable Transformer Encoder Neck

We propose the deformable transformer encoder neck used after the Res2net [13] and before the segmentation head. A detailed description of our network architecture is illustrated in Figure 5. Its purpose is to fuse features adaptively across multiple levels at learned locations so that the classification of each pixel considers the surrounding features with learnt weight. In this case, the attention to features inside the polyp can help the classification of pixels at ambiguous locations such as edges.

Deformable Transformer Encoder Neck (DTEN) The structure of DTEN is illustrated in Figure 6. Multi-scale feature maps $\boldsymbol{m}_l (l = 1, 2, 3)$ with resolutions $H_l \times W_l$ are passed to a convolutional layer to have the same number of channels and then are normalized to have an equal contribution. Then the feature maps are flattened and concatenated to form the input feature $\boldsymbol{m_f}$. The input feature along with the pre-generated reference points and the embedding statistics [44] are passed to the deformable encoder. The encoder outputs multi-scale enhanced feature maps $\boldsymbol{o}_l$ with resolutions the same as $\boldsymbol{m}_l$. For simplicity and sufficient details, only $\boldsymbol{o}_3$ which contains the finest features is utilized in the subsequent stacked Feature Add (FA) blocks.

Deformable Encoder The deformable encoder [44] enriches the input mainly by the deformable attention mechanism, as shown in Figure 7. It sums the selected features at learned sampling locations across multi-scales with learned attention weights. The detailed architecture of the encoder can be found in the supplementary material. The output of the encoder is then reshaped into the original resolutions, forming the enhanced multi-scale feature maps $\boldsymbol{o}_l$ as shown in Figure 6 for subsequent progressive feature compensation. A detailed explanation of the deformable encoder is in A.

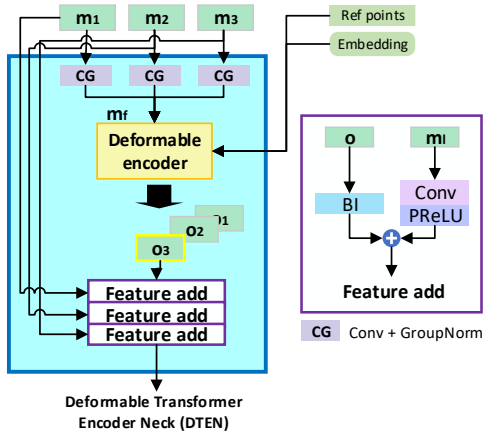


Figure 6. The detailed structure of the Deformable Transformer Encoder Neck (DTEN) as described in 3.2.5.

Feature Add (FA) Block The purpose of the FA block is to compensate the input feature map with enhanced features. The structure of a FA block is shown in Figure 6. It takes the enhanced feature map $\boldsymbol{o}$ and the original feature map $\boldsymbol{m_l}$ as inputs. The original feature map is embedded via a convolution layer and the PReLU. The enhanced map is interpolated to the same resolution as the original feature map. Three FA

Table 1. Ablation study with mDice and mIoU on five challenging datasets: ColorDB, ETIS, Kvasir, CVC-300 and ClinicDB. Upper part: the network is trained through our weak annotations. †: denotes models trained using fully-supervised training through regular dense annotations. The best results are in **bold**.

Method	ColorDB		ETIS		Kvasir		CVC-300		ClinicDB	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
L_p	0.327	0.263	0.218	0.168	0.555	0.488	0.240	0.174	0.479	0.448
L_{weak}	0.539	0.503	0.442	0.415	0.700	0.668	0.662	0.658	0.740	0.708
$L_{weak} + L_c$	0.604	0.544	0.501	0.442	0.730	0.677	0.729	0.678	0.771	0.718
L_{weak} + DTEN	0.609	0.538	0.541	0.472	0.728	0.665	0.754	0.702	0.772	0.707
L_{weak} +DTEN+ L_c	**0.667**	**0.588**	**0.596**	**0.517**	**0.768**	**0.709**	**0.795**	**0.728**	**0.807**	**0.746**
Backbone†	0.688	0.612	0.646	0.568	0.851	0.796	0.856	0.785	0.833	0.768
+DTEN†	**0.723**	**0.640**	**0.664**	**0.583**	**0.862**	**0.805**	**0.861**	**0.805**	**0.854**	**0.791**

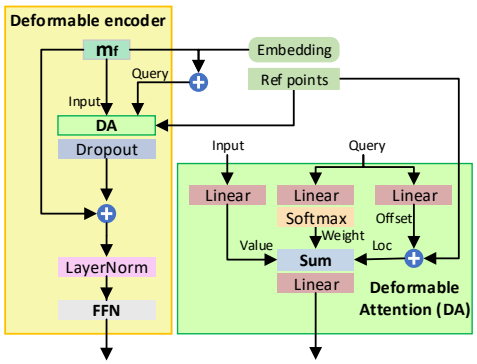


Figure 7. The structure of a single encoder in deformable vision transformers. The encoder enables the aggregation of useful features at learned locations with learnt significance across levels.

blocks are stacked to complement progressively the input features with the enhanced features by element-wise addition to output a more expressive feature map.

4. Experiments

4.1. Setup

Datasets and Evaluation Metrics We conduct experiments on five widely used polyp datasets, namely CVC-ColonDB [31], ETIS [30], Kvasir [17], CVC-T [35] and CVC-ClinicDB [3]. Kvasir contains 1,000 polyp images and CVC-ClinicDB contains 612 images from 31 colonoscopy clips. The composited training images come from these two datasets and the rest of them are used for testing. The other three testing datasets are totally unseen with challenging scenarios. We follow [11, 37] and employ two commonly used metrics, namely mean Dice (mDice) and mean IoU (mIoU), to evaluate the model performance for polyp segmentation.

Implementation Details Our model is implemented using Pytorch Toolbox [26] and trained on a GTX TITAN X GPU with a mini-batch size of 4. We adopt a 0.0005 weight decay for the Stochastic Gradient Descent (SGD) with a momentum of 0.9. For fair comparisons, both training and testing images are resized to 352×352, which is the same as the previous polyp segmentation methods.

Table 2. We substitute the edge loss in [42] with the proposed sparse foreground loss and apply the same SOD training and testing settings as [42] on three SOD evaluation metrics, namely F-measure [1] (F), E-measure [10] (E) and mean absolute error (M).

	Metric	Edge	L_f
ECSSD	F	0.862	0.854
	E	0.913	0.907
	M	0.063	0.063

4.2. Ablation Study

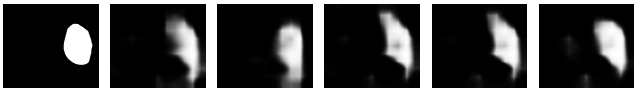
We conduct extensive experiments to analyze the merits of our proposed framework, WS-DefSegNet. Table 1 ablates our framework and shows that each component, namely L_{weak}, L_c, and DTEN, boosts the segmentation performance compared to training only using L_p [32].

4.2.1 Sparse Foreground Loss

As discussed in section 1, training a model using only the partial loss L_p causes a lot of false positives. Our proposed sparse foreground loss L_f addresses this problem as it is shown in Figure 2. Compared to Figure 2(c), Figure 2(d) shows more accurate segmentation maps, which are more similar to the ground truth with fewer false positives. The benefit of our sparse foreground loss is also reflected in the overall performance as shown in Table 1, providing gains

Table 3. Comparisons with different semi-supervised learning methods on five challenging datasets. L_{weak} refers to using the model after the weakly-supervised training without any semi-supervised training.

Method	ColorDB		ETIS		Kvasir		CVC-300		ClinicDB	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
L_{weak}	0.539	0.503	0.442	0.415	0.700	0.668	0.662	0.658	0.740	0.708
L_c	0.553	0.508	0.477	0.431	0.713	0.665	0.704	0.678	0.740	0.693
$L_{weak} + L_c(unweighted)$	0.559	0.513	0.483	0.439	0.716	0.668	0.702	0.667	0.748	0.701
$L_{weak} + L_c(weighted)$	0.604	0.544	0.501	0.442	0.730	0.677	0.729	0.678	0.771	0.718



(a) (b) (c) (d) (e) (f)

Figure 8. Difference between predictions of the two identical backbones with the same training settings and different semi-supervised methods. (a) Ground truth. (b) Predicted segmentation map of the first model. (c) Predicted segmentation map of the second model. (d) Only pseudo labels for semi-training. (e) L_{semi} without β for semi-training. (f) Ours.

of up to 42.2% and 48.4% in terms of mDice and mIOU respectively on CVC-300.

Additionally, in contrast to the previous work [42], which uses edge information and an auxiliary network to refine the segmentation maps, we do not use any extra information and networks. We simply use our sparse foreground loss to obtain more accurate segmentation maps and aid the model to localize objects.

In order to further demonstrate the effectiveness of our method, we conduct experiments on S-DUTS dataset [42] and substitute the edge loss in weakly Salient Object Detection (SOD) [42] with our sparse foreground loss as shown in Table 2. The results indicate that the proposed method can be exploited on other weakly-supervised tasks and can achieve similar performance.

4.2.2 Batch-wise Weighted Consistency Loss

We include the batch-wise weighted consistency loss L_c in the baseline L_{weak} in Table 1 for the semi-supervised training. The experimental results show that this method can increase the segmentation accuracy on both mDice and mIoU across all testing datasets. It can also be observed in Figure 2(e) that L_c eliminates the false positive pixels next to the polyp, and also improves the predicted segmentation maps compared with Figure 2(d).

Interestingly, the superiority of the batch-wise weighted consistency loss can be seen in Table 3. Apparently, without the aid of weights β_1 and β_2, L_{weak} contributes minor improvement compared to only training on pseudo labels during semi-supervised training. Using the batch-wise weighted consistency loss addresses the inconsistent issue caused by weak supervision (Figure 8(b) and (c)) by taking full advantage of the two predicted maps for semi-supervised training. When compared to using only pseudo-labels or using L_{semi} without weights β_1 and β_2, it can be seen from Figure 8(d), (e) and (f) our proposed solution provides a more accurate segmentation map with refined boundaries.

4.2.3 DTEN

To investigate whether the proposed DTEN benefits polyp segmentation, we compare the results with and without DTEN under different training regimes as shown in Table 1. Regarding the weakly- and semi-supervised training part, DTEN provides significant performance increase under all metrics and on all datasets when compared to using only our proposed loss functions. In the fully supervised training section, applying DTEN on top of the Res2net also enhances the performance. These results indicate the effectiveness of the proposed structure and demonstrate the importance of enhancing features for accurate segmentation.

4.3. Comparison with the state-of-the-arts

To further validate our proposed framework, we compare it with other state-of-the-art methods, namely, U-Net [28], U-Net++ [43], ResUNet++ [18], SFA [12], PraNet [11] and CAL [37] on five challenging polyp testing datasets. We directly report the results provided by each work. It should be noted that we are the only ones using weakly annotated images. Our results show that we can compete and even surpass methods that were trained in a fully supervised way as seen in Table 4. Also, we obtain competitive results compared to [37] which is the only other method that uses semi-supervised training. However, in contrast to our framework, [37] uses pixel-wise annotated images while we only use weakly-annotated images. Also, our method uses less than half of the averaged labeled pixels that [37] uses.

It is also worth noting that other state-of-the-art methods [12, 18, 28, 37, 43] may suffer from overfitting issues because they only obtain high performance on Kvasir and ClinicDB.

Table 4. Evaluation results of different methods on five datasets.*uses semi-supevised training. Ours: denotes our method that is trained using weakly- and semi-supervised training.

Method	Average Labeled Pixels	ColorDB		ETIS		Kvasir		CVC-300		ClinicDB	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net(MICCAI'15) [28]	13.4%	0.512	0.444	0.398	0.335	0.818	0.746	0.710	0.627	0.823	0.755
U-Net++(TMI'19) [43]	13.4%	0.483	0.410	0.401	0.344	0.821	0.743	0.707	0.624	0.794	0.729
ResUNet++(ISM'19) [18]	13.4%	-	-	-	-	0.813	0.793	-	-	0.796	0.796
SFA(MICCAI'19) [12]	13.4%	0.469	0.347	0.297	0.217	0.723	0.611	0.467	0.329	0.700	0.607
PraNet(MICCAI'20) [11]	13.4%	0.709	0.640	0.628	0.567	0.898	0.840	0.871	0.797	0.899	0.849
CAL(ICCV'21)* [37]	4.0%	-	-	-	-	0.810	0.716	-	-	0.893	0.826
Ours	1.9%	0.667	0.588	0.596	0.517	0.768	0.709	0.795	0.728	0.807	0.746

Table 5. Fine-tuning results with mDice and mIoU on five challenging datasets for different state-of-the-art approaches. †: denotes models trained using fully supervised training through regular dense annotations. The best results are in **bold**.

Method	ColorDB		ETIS		Kvasir		CVC-300		ClinicDB	
	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
Poolnet(pretrained) [23]	0.159	0.103	0.086	0.057	0.455	0.361	0.135	0.084	0.240	0.171
Poolnet†	0.439	0.403	0.330	0.327	0.774	**0.743**	0.543	0.528	0.629	0.622
Ours(L_{weak})	0.576	0.508	0.426	0.383	0.743	0.682	0.722	0.649	0.763	0.701
Ours($L_{weak} + L_c$)	**0.583**	**0.508**	**0.459**	**0.415**	**0.776**	0.708	**0.755**	**0.676**	**0.782**	**0.721**
A2dele(pretrained) [27]	0.219	0.153	0.225	0.161	0.470	0.352	0.359	0.271	0.287	0.195
A2dele†	0.450	0.461	0.378	0.406	**0.706**	**0.713**	0.666	0.718	0.588	0.633
Ours(L_{weak})	0.487	0.500	0.413	0.440	0.610	0.605	0.660	0.702	0.579	0.601
Ours($L_{weak} + L_c$)	**0.509**	**0.511**	**0.449**	**0.457**	0.662	0.645	**0.695**	**0.728**	**0.623**	**0.636**

Compared to them, ours achieves satisfactory performance on all five testing datasets. The results in Table 4 demonstrate the superior generalization ability of our framework.

4.4. Transfer Learning on Other Networks

In order to investigate the transferability of our method, we leverage our framework to adapt other networks that were trained on different tasks. First of all, we use two pre-trained SOD detectors, the RGB-trained Poolnet [23] and the RGB-D trained A2dele [27], and show that we can fine-tune them successfully using our novel loss functions L_{weak} and L_c as shown in Table 5. The baseline results show how each method performs without any adaptation. Impressively, simply fine-tuning both A2dele and Poolnet using our proposed loss functions outperforms fine-tuning in a fully supervised way. These results highlight the transfer learning ability of our framework and its potential to be used for different networks. Similarly to Table 1, it can be seen that each of our proposed loss functions provides a significant performance increase.

5. Conclusion

In this paper, we propose a novel framework WS-DefSegNet for weakly- and semi-supervised polyp segmentation. We create a weakly annotated polyp dataset (W-Polyp) by simply drawing sketches. This annotating method provides an efficient way for physicians to avoid manual labour.

We propose a sparse foreground loss that suppresses false positives. Furthermore, we propose a batch-wise weighted consistency loss to exploit two inconsistent segmentation maps caused by weak supervision during semi-supervised training. Also, we design a deformable transformer encoder neck (DTEN) as a way to enhance features before the segmentation head further improving performance.

Extensive experiments are conducted on five challenging datasets to demonstrate that each proposed component improves the segmentation accuracy and that our framework can even surpass the performance of some state-of-the-art methods trained in a fully supervised way.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009. 6

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2209–2218, 2019. 3

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 6

[4] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019. 3

[5] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 3

[6] John Scott Bridle, A.J.R. Heading, and David J. C. Mackay. Unsupervised classifiers, mutual information and 'phantom targets'. In *NIPS*, 1991. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 12

[8] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning (adaptive computation and machine learning). 2006. 3

[9] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 3

[10] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 6

[11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 1, 2, 3, 6, 7, 8, 11

[12] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 302–310. Springer, 2019. 1, 2, 7, 8

[13] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652–662, 2019. 5

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. 1

[15] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging*, 38(10):2281–2292, 2019. 2

[16] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019. 3

[17] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020. 6

[18] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019. 2, 7, 8

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 3

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 3

[21] Zhuoying Li, Junquan Pan, Huisi Wu, Zhenkun Wen, and Jing Qin. Memory-efficient automatic kidney and tumor segmentation based on non-local context guided 3d u-net. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 197–206. Springer, 2020. 2

[22] Ailiang Lin, Jiayu Xu, Jinxing Li, and Guangming Lu. Contrans: Improving transformer with convolutional attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 297–307. Springer, 2022. 3

[23] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3912–3921. IEEE, 2019. 8

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[25] Yanglan Ou, Ye Yuan, Xiaolei Huang, Stephen TC Wong, John Volpi, James Z Wang, and Kelvin Wong. Patcher: Patch transformers with mixture of experts for precise medical image segmentation. *arXiv preprint arXiv:2206.01741*, 2022. 3

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 6

[27] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9060–9069, 2020. 8

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 7, 8

[29] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 3

[30] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014. 6

[31] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 6

[32] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018. 2, 6

[33] Youbao Tang, Ning Zhang, Yirui Wang, Shenghua He, Mei Han, Jing Xiao, and Ruei-Sung Lin. Accurate and robust lesion recist diameter prediction and segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 535–544. Springer, 2022. 3

[34] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3, 4

[35] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017, 2017. 6

[36] Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, and Xiaopeng Zhang. Da-net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2022. 3

[37] Huisi Wu, Guilian Chen, Zhenkun Wen, and Jing Qin. Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3489–3498, 2021. 1, 3, 6, 7, 8

[38] Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022. 3

[39] Jingshan Xu, Chuanwei Zhou, Zhen Cui, Chunyan Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, and Jian Yang. Scribble-supervised semantic segmentation inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15354–15363, 2021. 2

[40] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3234–3242, 2021. 2, 3

[41] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 3

[42] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12546–12555, 2020. 2, 3, 4, 6, 7

[43] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 2, 7, 8

[44] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2021. 3, 5, 11