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

Guangyu Ren, Michalis Lazarou, Jing Yuan, Tania Stathaki

Imperial College London

Motivation and contributions

Motivation:

- Polyp segmentation is crucial for diagnosis of colorectal cancer.
- Annotated datasets are tedious and time-consuming to produce.
- Save physicians' valuable time.

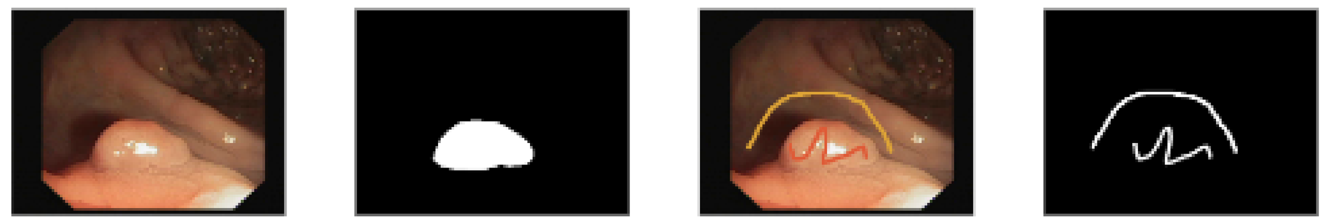
Contributions:

- Provide the first weakly-annotated polyp dataset, **W-Polyp**.
- Provide the first weakly- and semi-supervised training framework, **WS-DefSegNet**.
- Propose a novel weakly-supervised loss function, **Sparse foreground loss**.
- Propose a novel progressive multi-scale architecture with a self-attention mechanism, **DTEN**.

W-Polyp dataset

Information:

- 1450 images in total.
- 750 images weakly annotated with simple sketches, lines, scribbles and circles.
- 700 images left unlabeled.

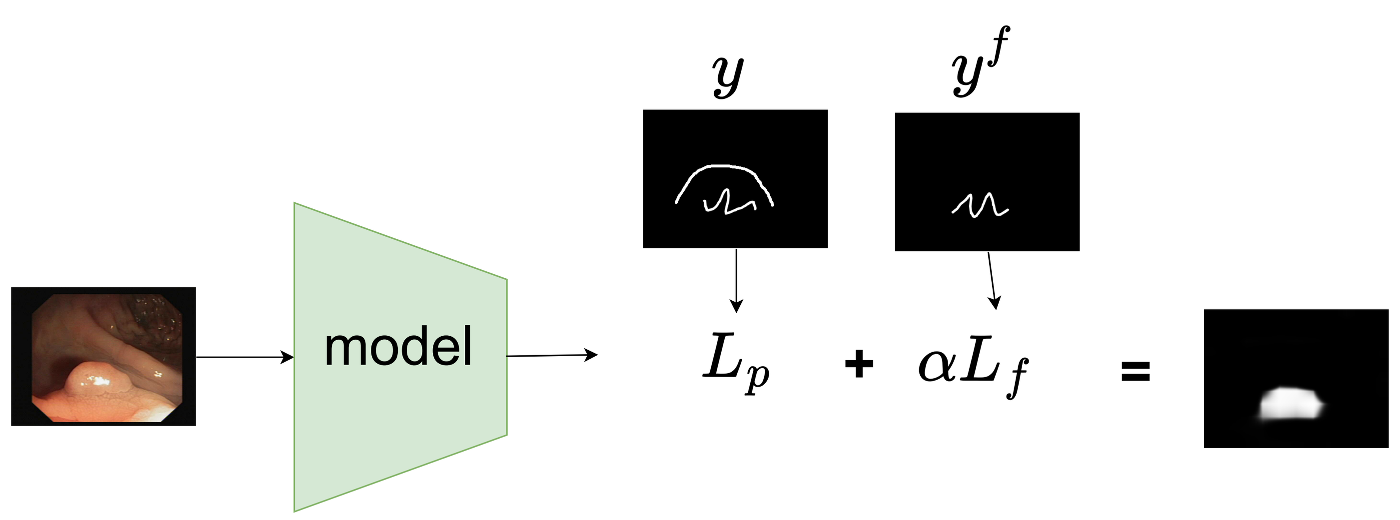


(a) (b) (c) (d)

Visualization of weak annotations. (a) RGB image. (b) Original ground truth. (c) Foreground and background annotations. (d) Our weak annotations.

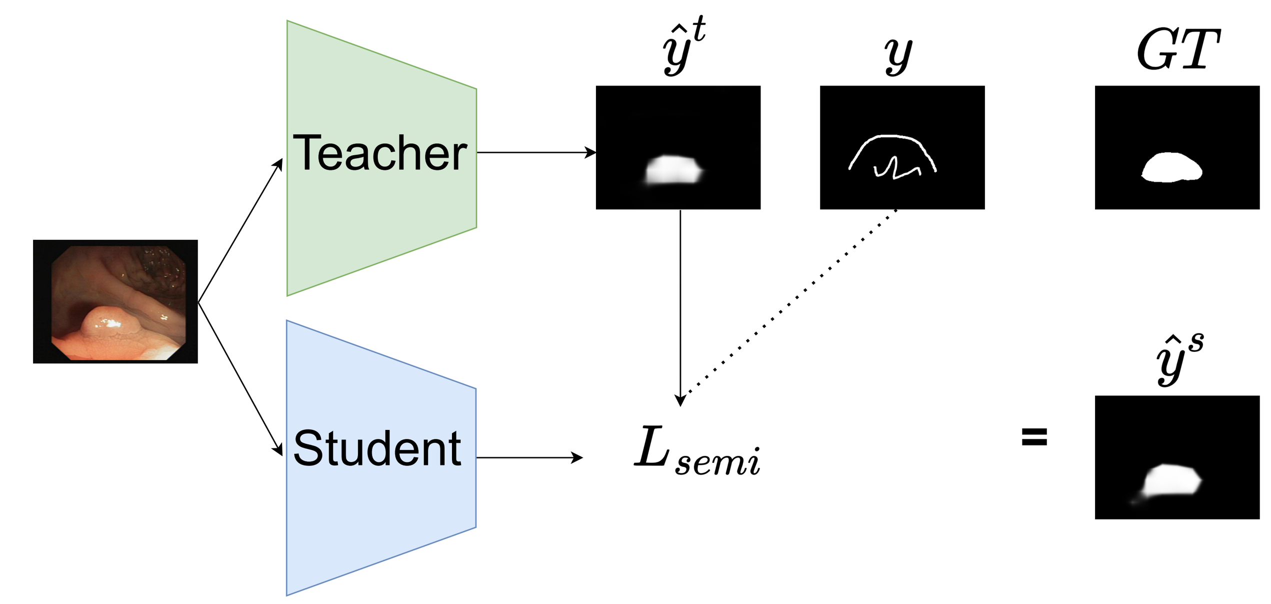
Weakly-supervised training

- Weighted loss between two loss functions.
- Partial cross entropy loss, L_p, utilizes information for both background and foreground annotations.
- Sparse foreground loss, L_f, utilizes information of only the foreground annotations.

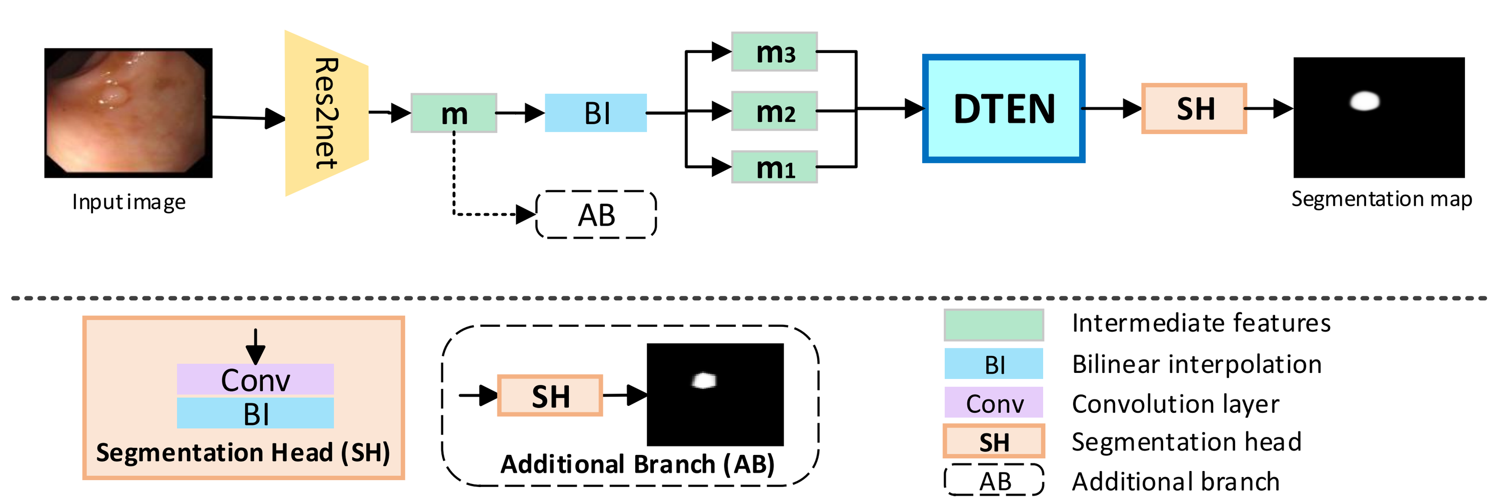


Semi-supervised training

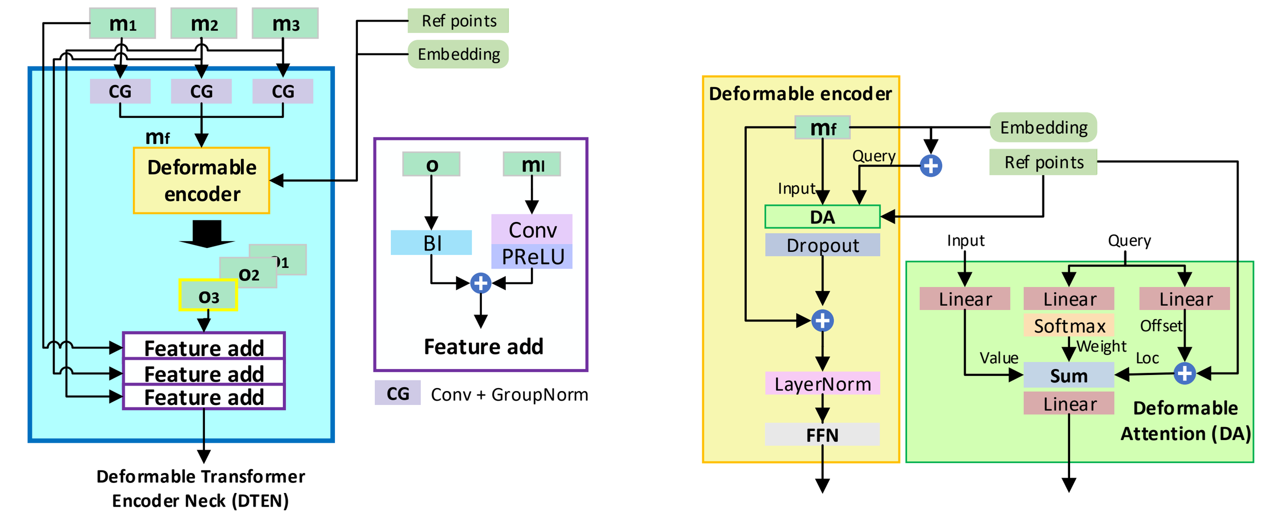
- Utilize a teacher-student training paradigm.
- Teacher model provides predicted pseudo-labels that are used along with the weakly-annotated labels to train the student model.



Architecture design



- The Deformable Transformer Encoder Neck (DTEN) fuses features adaptively across multiple levels at learned locations with learned weights so that the classification of each pixel considers the surrounding features. This helps the classification of pixels at ambiguous locations such as edges.
- The Feature Add (FA) blocks progressively compensate the input feature map with enhanced features to produce more expressive feature maps.



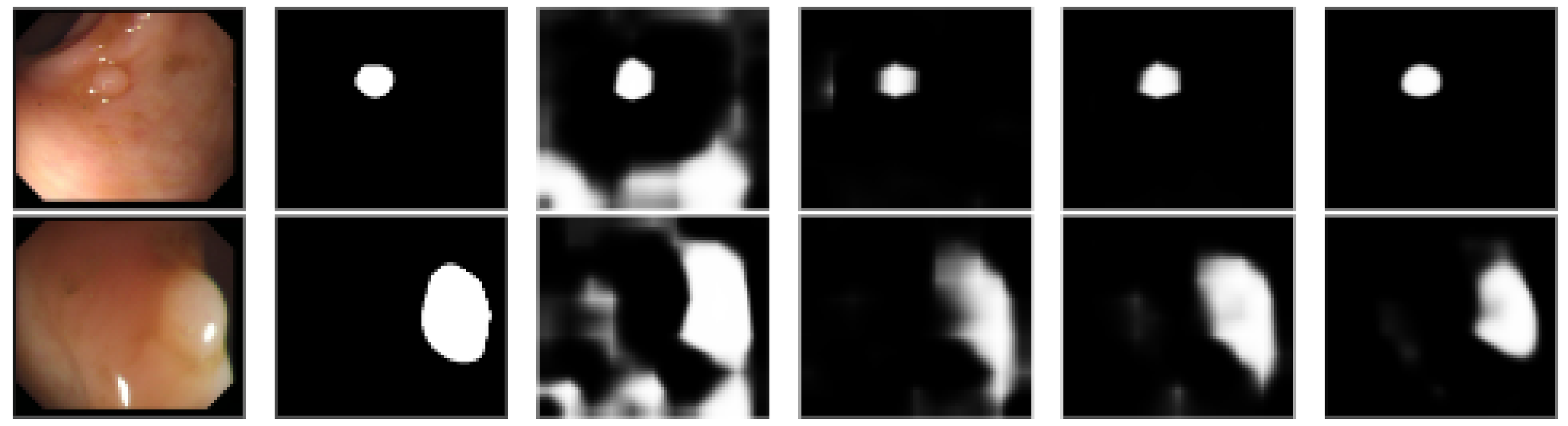
Implementation details

- Our implementation is based on PyTorch and OpenCV.

Ablation study

Method	ColonDB		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**

Ablation study with mDice and mIoU on five challenging datasets: ColonDB, 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**.



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

Visual comparison of ablation study. (a) RGB image. (b) Original ground truth (c) L_p. (d) $L_p + \alpha L_f$. (e) L_{semi}. (f) +DTEN.

State of the art comparisons

Method	Average Labeled Pixels	ColonDB		ETIS		Kvasir		CVC-300		ClinicDB	
		mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU	mDice	mIoU
U-Net(MICCAI'15)[4]	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)[6]	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)[3]	13.4%	-	-	-	-	0.813	0.793	-	-	0.796	0.796
SFA(MICCAI'19)[2]	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)[1]	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)*[5]	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

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.

[1] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. 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.

[2] Y. Fang, C. Chen, Y. Yuan, and K.-y. 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.

[3] D. Jha, P. H. Smedsrud, M. A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, and H. D. Johansen. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255. IEEE, 2019.

[4] O. Ronneberger, P. Fischer, and T. 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.

[5] H. Wu, G. Chen, Z. Wen, and J. 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.

[6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. 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.

Contact: {g.ren19, michalis.lazarou14, j.yuan20, t.stathaki}@imperial.ac.uk