Heterogeneous model ensemble for automatic polyp segmentation in endoscopic video sequences

Thuy Nuong Tran[1], Fabian Isensee[2,3], Lars Krämer[2,3], Amine Yamlahi[1], Tim Adler[1], Patrick Godau[1], Minu Tizabi[1] and Lena Maier-Hein[1]

[1] *Div. Intelligent Medical Systems, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[2] *Div. Medical Image Computing, DKFZ, Heidelberg, Germany*
[3] *Applied Computer Vision Lab, Helmholtz Imaging*

Abstract

The detection and segmentation of polyps during colonoscopy can substantially contribute to the prevention of colon cancer. Assisting clinicians using automated systems can mitigate the risk of human error. In this work, we present our polyp segmentation approach, submitted to the EndoCV2022 challenge. Common polyp segmentation methods are based on single-model, single-frame predictions. This work presents a symbiosis of three separate models, each with their own strength, as part of a segmentation pipeline and a post-processing step designed to leverage unique predictions for more temporally coherent results.

Keywords

Polyp segmentation, Temporal coherence, High resolution, Heterogeneous ensemble

1. Introduction

Colorectal cancer is one of the most commonly found cancer types, ranking second in females and third in males [1]. By detecting and subsequently resecting polyps during colonoscopy screenings, the risk of developing the disease can be reduced significantly. With the advance of machine learning in the medical domain, deep learning-based methods have the potential to assist in detecting and segmenting these polyps with high accuracy. The EndoCV2022 challenge[2] addresses generalizability of such deep learning models for segmentation in endoscopic video sequences. The method presented in this paper tackles this issue with three primary design decisions: (1) The provided challenge dataset underwent a curation process that ensures annotation quality. (2) An ensemble of three networks with complementary strengths was trained for the segmentation prediction. (3) Finally, a post-processing step was implemented to address false-negative frames caused by majority vote. A fallback mechanism was set to reweight the predictions of a single model in order to enable unique predictions.

2. Datasets

The dataset provided by the EndoCV2022 polyp segmentation sub-challenge[2, 3, 4] consists of 46 sequences of varied length, totalling 3,290 image frames and their corresponding polyp segmentation masks. Furthermore, three public polyp segmentation datasets were added as external data, namely CVC-ColonDB[5], CVC-ClinicDB[6] and ETIS-Larib[7], to enrich the diversity of the dataset. These account for 1,108 additional training images, resulting in 4,398 frames in total.

3. Methodology

Our challenge strategy rests on three main pillars: (1) A data pre-processing step to ensure high data annotation quality, (2) the network architecture selection and training step, which yields the segmentation models, and (3) a post-processing step, which leverages model heterogeneity and uses structural similarity[8] of consecutive frames in order to handle false-negative masks. An overview is depicted in Fig.1.

3.1. Data pre-processing

Correct data annotation of the training set is crucial to the learning capabilities of any segmentation model. In order to ensure annotation quality, the provided challenge dataset was curated by manually removing images with implausible or temporally inconsistent annotations, to the best of our judgement. An example is shown in Fig.2. This was conducted under the assumption that false annotation would harm the training process more

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

t.tran@dkfz-heidelberg.de (T. N. Tran)

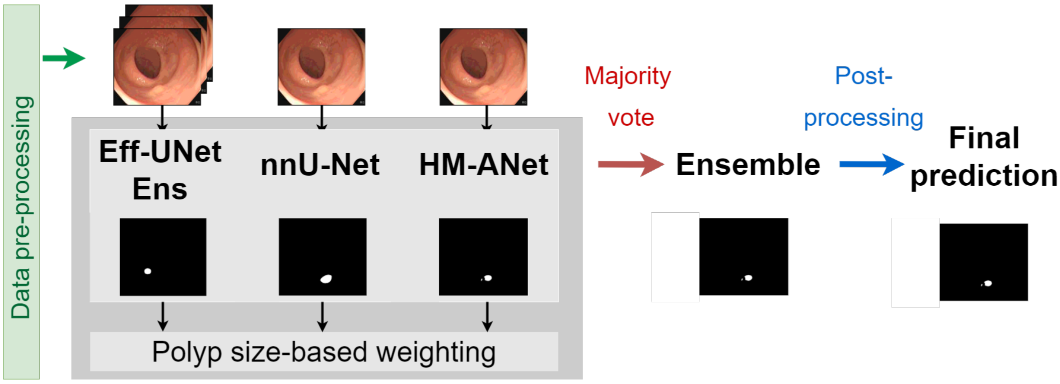

Figure 1: Overview of heterogeneous model ensemble pipeline. Data is curated. Predictions of Efficient-UNet[9] ensemble, nnU-Net[10] and Hierarchical Multi-Scale Attention Network[11] are combined. Post-processing yields final prediction.

than having a larger number of frames for training. The external datasets described in section 2 underwent the same selection process. Including external data, the resulting training dataset amounted to 4,106 image-mask pairs.

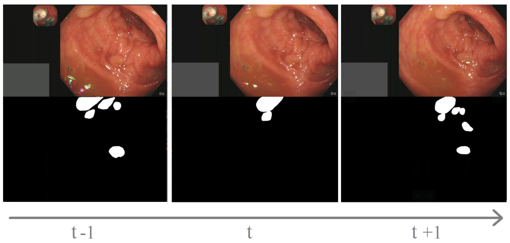

Figure 2: Example of inconsistent annotation. The upper row depicts three consecutive frames of the provided *seq23_endocv22* sequence. The lower row shows the segmentation mask. The image at position t has fewer polyps annotated compared to the neighboring frames, despite the polyps not being obstructed or out of sight.

3.2. Neural network architectures

In order to solve the polyp segmentation task, a model ensemble was designed that consists of parts with complementary strengths. This was realized by using an nnU-Net[10], which is configured to automatically adapt its pre-processing and training framework to different datasets, and thus serves as a strong segmentation base, a Hierarchical Multi-Scale Attention Network[11], which combines predictions of multiple scales for a better prediction performance, and an ensemble of Efficient-UNets[9], one of which is equipped with an internal GRU-layer to process temporal information. By focusing on incorporating temporal as well as high-resolution information, we expected more knowledge to be leveraged from the provided high-resolution video sequences.

3.2.1. nnU-Net

The nnU-Net is able to automatically determine key decisions to set up the segmentation pipeline for training, irrespective of the dataset. While it has ranked first on many 3D-segmentation challenges[1], its self-configuring strategy can also be applied to 2D images. The nnU-Net was expected to provide a solid base prediction.

3.2.2. Hierarchical Multi-Scale Network

By treating the polyp segmentation as a classic computer vision task, it is possible to use established segmentation models that perform well on complex natural images. The Hierarchical Multi-Scale Attention Network (HM-ANet) was chosen as it is a state-of-the-art architecture in semantic segmentation on Cityscapes[2]. The HM-ANet operates on higher resolutions and combines predictions from different scales. This was expected to result in a precise polyp segmentation, irrespective of the size of the polyp.

3.3. Efficient-UNet Ensemble

Most of the current segmentation models operate on a frame-by-frame basis. In order to capture temporal information, one approach is to incorporate a recurrent neural network layer to a standard segmentation model, such as a Gated Recurrent Unit (GRU)-layer. The chosen base segmentation model is the Efficient-UNet (Eff-UNet). It is an encoder-decoder architecture with an EfficientNet as its backbone, which is able to scale with model size

[1] medicaldecathlon.com,https://kits19.grand-challenge.org,https://www.med.upenn.edu/cbica/brats2020

[2] www.cityscapes-dataset.com

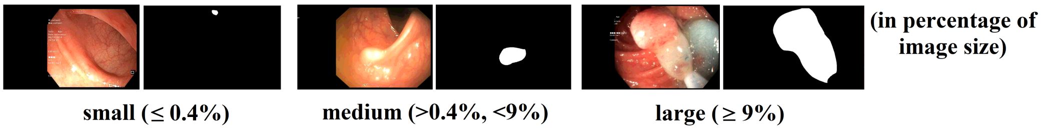

Figure 3: Example of differently sized polyp images with their segmentation masks from the provided challenge set.

and outperform other ConvNet backbones. One GRU-layer was added to the bottleneck of the Eff-UNet, to form an Eff-GRUNet. Consecutive images are loaded in batches of size two. They are encoded, pooled, flattened, sequentially fed into the GRU-layer and then reshaped and fed to the decoder. The Eff-UNet is trained separately from the Eff-GRUNet. Variants of both combined form the Eff-UNet ensemble.

3.3.1. Combining networks and weighting

Since the HM-ANet operates on high resolutions, it was expected that it performs well on very small polyps, as well as being able to fully capture larger polyps in their entirety. During ensembling, the HM-ANet was designed to be weighted higher for the small and large polyps. Since there is no standardized definition of polyp sizes, the thresholds were set empirically by observing reference labels of public polyp datasets[5, 6, 7]. An example is shown in Fig.3.

3.4. Post-processing by reweighting

To mitigate the error of false-negative predictions, a post-processing step is added that considers empty segmentation masks and their surrounding frames. If a neighboring frame is polyp-positive and is similar to the current frame, then any non-empty prediction of the current frame is reweighted, effectively allowing a polyp-positive prediction despite non-majority. The similarity score used for this approach is the structural similarity score(SSIM), as it is able to take texture into account.

4. Experiments and Results

The original training dataset was split into four parts using GroupK-fold for 4-fold cross-validation(CV) training, balancing the number of frames and sequence IDs. Each fold has 11-12 sequences with around 750 frames. The following subsections describe the implementation details and experiment results after hyperparameter optimization.

4.1. Implementation details

The nnU-Net was used as a framework and manual changes were made to its automatically generated configuration. The short edge of the image was resized to 512px with the other being resized according to aspect ratio. The patch-size was set to 448 x 448. The data was then heavily augmented with operations such as rotation, intensity and gamma augment, scaling, mirroring and blurring.

For the HM-ANet, the data was normalized and random scaling between [0.5,1], random crop to 512x1024, RGB-shift, and random vertical and horizontal flipping was performed. The model was initiated with weights pre-trained on PaddleClas[3]. The training was conducted in three phases: 1) Training the model on original challenge data, 2) fine-tuning the model on challenge and external data, and 3) fine-tuning again on challenge data only.

For the Eff-UNet ensemble, the data was resized to 480x480 (Eff-UNet_480) and 256x256 (Eff-UNet_256), incorporating different resolutions. Resizing to 256x256 was chosen for the Eff-GRUNet, to fit memory restrictions. Augmentations such as rotation, elastic and grid deformation were used.

In order to combine the predictions, the segmented polyps were divided into small ($\leq$ 0.4% of image size), large ($\geq$9% of image size) and medium (rest) polyps. If polyps were predicted as small or large, the weight of the HM-ANet was increased to 0.5, while the others were decreased to 0.25 each. If the polyp was of medium size, the models were weighted equally at 0.33. The final segmentation was formed by thresholding the weighted predictions at 0.5. To address false-negatives resulting from an unmet majority criterion, unique single-model predictions were encouraged if neighboring images were structurally similar (SSIM $>$ 0.9) and predicted to be polyp-positive. The single model prediction weight was then increased to 0.5. This proved to solve some false-negative cases, as illustrated in Fig.4.

4.2. Single model experiment results

All final single model DSC scores are reported in Table 1. The nnU-Net was trained with external data added to the

[3] paddleclas.readthedocs.io/en/latest/index.html

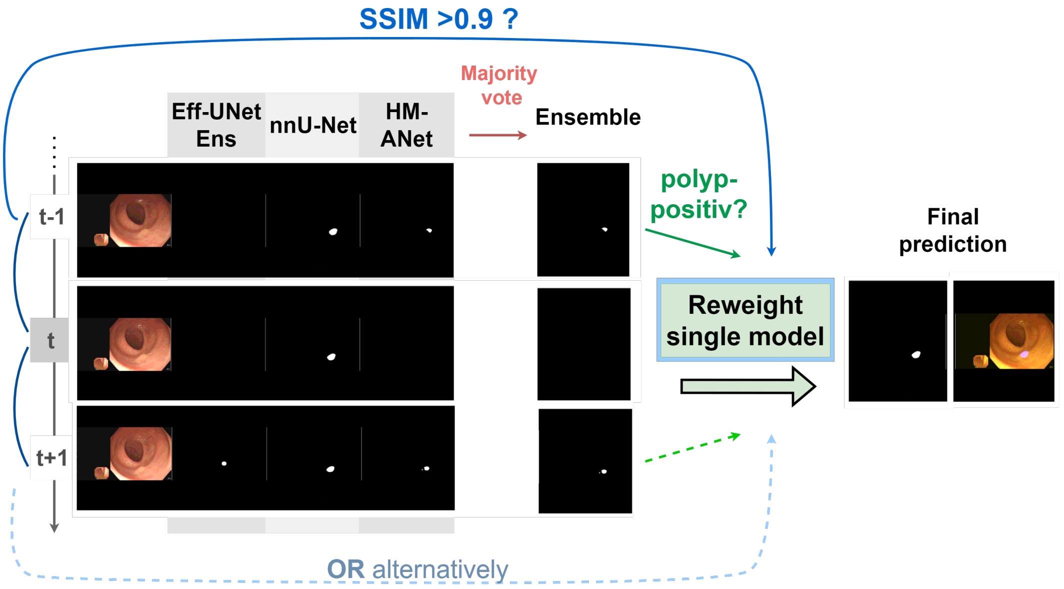

Figure 4: Example of post-process reweighting. The ensemble prediction at time step t is empty. Because SSIM > 0.9 and prediction is non-empty for at least one of the neighboring images, the single model prediction is weighted with 0.5.

training set, resulting in a mean CV DSC score of 0.67. Training only on the challenge set or the external dataset resulted in a worse DSC score of 0.57 and 0.55.

The HM-ANet had a mean CV DSC score across all folds of 0.70. During training and inference, predictions of scales [0.5,1] were combined. Experiments with scales of [0.5,1,2] resulted in a worse performance of 0.69 with more false-positives in empty images. Training in three steps as described in sub-subsection 4.1. yielded the best result. Other training strategies such as training on a combined dataset or pre-training on the external dataset and fine-tuning on the official dataset resulted in a worse performance. 4-fold cross-validation was used to determine the stopping epochs for all three phases. A final inference model was then trained on the entire dataset.

The three Eff-UNet models were each trained on the combined dataset over four folds, resulting in 12 models. The mean CV DSC scores of the Eff-UNet_480, Eff-UNet_256 and Eff-GRUNet were 0.69, 0.71 and 0.62, respectively. As an alternative experiment, the Eff-UNet_480 was trained with external data for pre-training and challenge data for fine-tuning. This performed worse compared to using the combined dataset, resulting in a mean CV DSC score of 0.65. In order to decrease inference time, two Eff-UNet_480, one Eff-UNet_256 and one Eff-GRUNet were selected for the ensemble, based on validation score and fold representation. The final prediction was determined by majority vote. The mean CV DSC score of the final ensemble was 0.70.

Table 1

Cross-Validation scores of all models, including the components of the Eff-UNet ensemble. Underscored values indicate selection for Eff-UNet ensemble. Bold values indicate components of the final heterogeneous ensemble.

DSC score	*Fold 0*	*Fold 1*	*Fold 2*	*Fold 3*	*Mean*
nnU-Net	0.65	0.84	0.70	0.50	**0.67**
HM-ANet	0.67	0.82	0.69	0.60	**0.70**
Eff-UNet_480	0.67	0.80	0.69	0.62	0.69
Eff-UNet_256	0.68	0.80	0.71	0.65	0.71
Eff-GRUNet	0.61	0.72	0.58	0.60	0.62
Eff-UNet Ens	0.67	0.80	0.71	0.60	**0.70**

4.3. Reweighting and ensembling results

In order to test the reweighting strategy of the HM-ANet, the proportion of small and big polyps was calculated for the validation splits. For folds 0-3, the ratios were 45%, 28%, 37%, and 65%. For single models, fold 1 had the most medium polyps and highest average CV score. Fold 3 has the most non-medium polyps, and the lowest average CV score. However, the difference in DSC scores between models is small. Since the ratio was highest for fold 3, an experiment is conducted where the three single models were validated on only the small and big polyp images of fold 3 (n = 483 out of 738 frames). The resulting DSC scores are 0.63, 0.66 and 0.70. The simple ensemble receives a score of 0.73 and the ensemble with reweighting of HM-ANet a score of 0.74. Adding post-processing did not decrease or increase the score for this validation set.

5. Conclusion

Our investigation showed that the HM-ANet was favorable for small and large polyp cases, which our dedicated weighing strategy takes into account during ensembling. Notably, on a dataset with small and big polyps, it achieves a DSC score of 0.74, improving the best-performing single model HM-ANet by 0.04. The post-processing leverages self-adaptive training as well as temporal and high resolution information by enabling unique predictions of all three heterogeneous components, resulting in less false-negative predictions. The inference time as the sum of the slowest component (nnU-Net) and the ensembling step is 0.71 fps.

6. Compliance with ethical standards

This work was conducted using public datasets of human subject data made available by [2, 3, 4, 5, 6, 7].

7. Acknowledgments

This project was supported by a Twinning Grant of the German Cancer Research Center(DKFZ) and the Robert Bosch Center for Tumor Diseases(RBCT). Part of this work was funded by Helmholtz Imaging(HI), a platform of the Helmholtz Incubator on Information and Data Science.

References

[1] F. A. Haggar, R. P. Boushey, Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors, Clinics in colon and rectal surgery 22 (2009) 191–197.

[2] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022). doi:10.48550/arXiv.2202.12031.

[3] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, et al., Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical image analysis 70 (2021) 102002. doi:10.1016/j.media.2021.102002.

[4] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021). doi:10.48550/arXiv.2106.04463.

[5] J. Bernal et al., Towards automatic polyp detection with a polyp appearance model, Pattern Recognition 45 (2012) 3166–3182.

[6] J. Bernal et al., Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Computerized Medical Imaging and Graphics 43 (2015) 99–111.

[7] J. Silva, A. Histace, O. Romain, X. Dray, B. Granado, Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer, International journal of computer assisted radiology and surgery 9 (2014) 283–293.

[8] Z. Wang et al., Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing 13 (2004) 600–612.

[9] B. Baheti et al., Eff-unet: A novel architecture for semantic segmentation in unstructured environment, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 358–359.

[10] F. Isensee et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature methods 18 (2021) 203–211.

[11] A. Tao et al., Hierarchical multi-scale attention for semantic segmentation, arXiv preprint arXiv:2005.10821 (2020).