Polyp Detection in Colonoscopy Images using Deep Learning and Bootstrap Aggregation

Gorkem Polat[a], Ece Isik-Polat[a], Kerem Kayabay[a] and Alptekin Temizel[a,b]

[a]*Graduate School of Informatics, Middle East Technical University, Ankara, Turkey*
[b]*Neuroscience and Neurotechnology Center of Excellence, Ankara, Turkey*

Abstract

Computer-aided polyp detection is playing an increasingly more important role in the colonoscopy procedure. Although many methods have been proposed to tackle the polyp detection problem, their out-of-distribution test results, which is an important indicator of their clinical readiness, are not demonstrated. In this study, we propose an ensemble-based polyp detection pipeline for detecting polyps in colonoscopy images. We train various models from EfficientDet family on both the EndoCV2021 and the Kvasir-SEG datasets, and evaluate their performances on these datasets both in- and out-of-distribution manner. The proposed architecture works in near real-time due to the efficiency of the EfficientDet architectures even when used in an ensemble setting.

Keywords

Polyp detection, Colonoscopy, Medical image Processing, Deep Learning

1. Introduction

Recent studies show that colorectal cancer is the third common cancer type with the second-highest mortality rate [1, 2]. It is estimated that there were more than 1.9 million new diagnoses of colorectal cancer and 935.000 deaths in 2020. It accounts for approximately one-tenth of all cancer cases and deaths [1]. Colonoscopy is the most common screening test for early-stage detection of colorectal cancer incidences and removal of polyps and adenomas, and potentially reducing mortalities [3, 4, 5]. However, effective polyp detection depends on the gastroenterologist's practical skills, and it is reported that approximately 20% of polyps may be missed [6]. Therefore, automatic and accurate detection of polyps is an essential aid to medical practice [6, 7].

Although there are many studies utilizing advanced machine learning methods on polyp detection, it has been shown that deep learning models could overfit, resulting in institutional biases [8]. While these models show good performance against test data from the same institution, they may not generalize well to external data from different institutions. In this work, we propose an ensemble-based polyp detection architecture for detecting polyps in colonoscopy

images. We evaluate the detection performance on both EndoCV2021 and Kvasir-SEG datasets. The results show that the proposed architecture improves the detection performance compared to the baseline methods. In addition, it has favorable out-of-distribution test results and near real-time processing speed.

The rest of the paper is organized as follows. The related work in the literature is reviewed in Section 2. In Section 3, the proposed approach and the details of the methodology are presented. The experimental design and training details are given in Section 4. The experiment results and discussion are given in Section 5. Finally, the conclusion of this study is given in Section 6.

2. Related Work

Automatic detection of polyps is an extensively studied subject. In earlier studies, proposed methods utilize machine learning methods and mainly focus on the shape, texture, or color properties of the polyps [9, 10, 11, 12]. The method proposed in [13] uses an ensemble architecture combining the results of convolutional neural network (CNN) models specialized for each polyp feature (color, texture, shape, and temporal information). Various approaches and methodologies, proposed in the context of the Automatic Polyp Detection Challenge [14], based on handcrafted features, end-to-end learning using CNNs, and their combinations are reported in [15]. In [16], a region-based CNN (i.e., Inception ResNet) is trained on both colonoscopy images and videos for the polyp detection task. Also, data augmentation methods such as rotating, scaling, shearing, blurring, and brightening were applied to increase the number of training samples. In [17], a multi-threaded deep learning algorithm that uses a smaller receptive field focusing on local features has been proposed for polyp detection using a large number of data with varying morphology. In [18], various methods that use deep learning and hand-crafted global features have been proposed to perform pixel-based, frame-based, and block-based segmentation and detection. In [19], state-of-the-art deep learning methods for polyp detection, localization, and segmentation were compared on the basis of both accuracy and speed with extensive experiments on the Kvasir-SEG [20] dataset and a new method, ColonSegNet, was proposed. This method is based on an encoder-decoder architecture for segmentation and provides promising performance results in real-time, in comparison to the state-of-the-art methods.

Studies in the literature generally focus on single model performance evaluated on the same distribution dataset. In this study, we propose a bagging-type ensemble-based method and evaluate the proposed architecture on datasets coming from different institutions to investigate its generalizability potential.

3. Proposed Method

3.1. Base Models

In this study, we use EfficientDet [21] model family due to their prominent speed and detection performance. EfficientDet models use EfficientNet [22] as backbone, weighted bi-directional feature pyramid network (BiFPN) for multiscale feature fusion, and shared class/box prediction
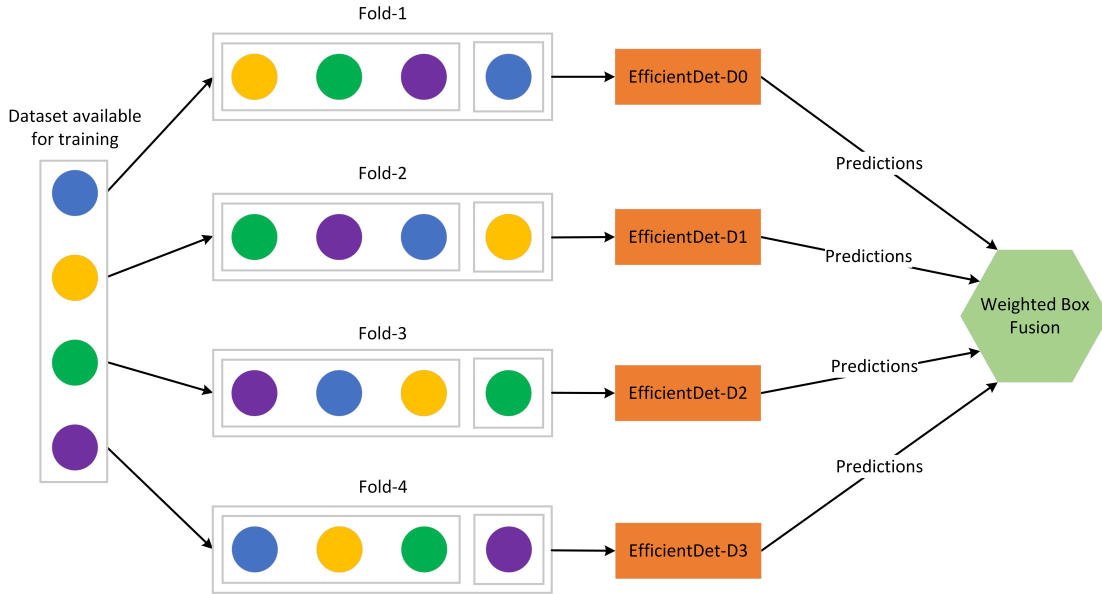
Figure 1: Flowchart of the proposed approach.

networks for bounding box classification and regression. The compound scaling method of the model scales the resolution, depth, and width for backbone, feature network, and box/class prediction networks at the same time. In this study, we have experimented with D0, D1, D2, and D3 versions and also evaluated their ensemble using bagging as described in the next section.

3.2. Ensemble of Models

Bootstrap aggregating (bagging) [23] is a type of ensemble method. The key idea of the bagging is to aggregate multiple versions of the predictors, which are trained on bootstrap replicates of the training set. Although bootstrap samples are generally formed by randomly picking data points with replacement, after reserving a part of the dataset for the test set, we split the rest into four cross-validation folds to generate folds as different from each other as possible. Each model is trained on a different fold and uses 75% of the samples in the fold for training and 25% for validation. When individual models are trained on a fixed training-validation split, the validation set is not seen by any model, which restricts each model to its respective training set. On the contrary, bagging-style ensemble provides that the combination of models trained on different folds covers the whole training dataset.

Bounding boxes generated individually by each model are then merged using weighted box fusion [24]. Unlike previous fusion techniques, which are generally based on keeping the highest confidence box and removing others, used in this domain [14, 25], weighted box fusion constructs an averaged box utilizing the confidence scores of the predicted boxes. In the previous approaches (NMS [26], soft-NMS [27]), the highest confidence boxes may not be the best rectangle for the ground truth object and the quality of the fused box is significantly improved with the weighted box fusion [24]. The overall architecture of the proposed approach

is given in Figure 1.

4. Experimental Design

In order to evaluate the performance of the individual models and compare them with the proposed bagging ensemble architecture, we designed two sets of experiments. First, the models are trained on the EndoCV2021 dataset and results are reported on both the EndoCV2021 test set, samples of which are from the same distribution, and Kvasir-SEG, samples of which are from another institution. Then the same models are trained on the Kvasir-SEG and results are reported on the Kvasir-SEG test set as well as the whole EndoCV2021.

The EndoCV2021 dataset [28] consists of 1449 images from 5 different centers (center-1: 256, center-2: 301, center-3: 457, center-4: 227, center-5: 208). The resolution of the images varies between 572×498 and 1920×1080. In addition, organizers provided 1793 frames from 15 different video segments. Since subsequent frames are very similar, we only sampled 75 frames out of 1793 frames to prevent bias towards a video segment due to over-representation. In total, 1524 images were split into training, validation, and test sets as 70%, 15%, and 15%, respectively. The validation set is used for hyperparameter tuning and the results are reported using the withheld test set. Ensemble of models on the fixed training set is also calculated for comparison with bagging ensemble. For the bagging ensemble, training and validation sets, which corresponds to 85% of the overall dataset, are combined and cross-validation folds are formed within it. Kvasir-SEG dataset contains 1000 polyp images and their corresponding ground-truth. The resolution of the images varies between 332×487 and 1920×1072. All 1000 images are used as an out-of-distribution test set. Results are given in Table 1.

In addition, we compared the proposed architecture's performance by training on the Kvasir-SEG. Kvasir-SEG dataset was split into training and test sets as 80% and 20%, respectively. The training dataset, then, was split into four cross-validation folds for the training of individual models. The proposed architecture has been evaluated on the folds, in the same way it was done for the EndoCV2021 dataset. Performances of two bagging ensembles, which are trained on the Kvasir-SEG and EndoCV2021 datasets, on the Kvasir-SEG test set are given in Table 2. We also evaluated the performances of the single EfficientDet models and bagging ensemble model trained on the the Kvasir-SEG dataset on the whole EndoCV2021 dataset to compare cross-dataset performance.

In order to increase the variance and for better generalization performance, the following data augmentation techniques have been used: Scale jittering (0.2, 2.0), horizontal flipping, and rotating (0°-360°). During the experiments, it was observed that brightness and contrast augmentations have a detrimental effect on the performance and, hence, they have been excluded. Since original images have varying resolutions, the maximum dimension of each image (width or height) is rescaled to the input size of the respective EfficientDet version, then, smaller dimension is rescaled with the same rate to preserve the aspect ratio. Since EfficientDet models require square inputs, zero padding is applied.

Models have been optimized by Adam optimizer [29]. Learning rate scheduling has been used, which decreased the learning rate by a factor of 0.2 whenever validation set loss did not decrease for the previous ten epochs. Early stopping has been used to terminate the training

Table 1

Detection results for the models trained on EndoCV2021. Models were tested both on EndoCV2021 test set of round II and Kvasir-SEG. Individual EfficientDet model results are given for the same training-validation split.

Model	EndoCV2021 Test Set			Kvasir-SEG as Test Set		
	mAP@50	mAP@75	mAP	mAP@50	mAP@75	mAP
EfficientDet-D0	0.783	0.567	0.509	0.859	0.646	0.568
EfficientDet-D1	0.772	0.559	0.496	0.862	0.673	0.574
EfficientDet-D2	0.784	0.592	0.527	0.858	0.668	0.588
EfficientDet-D3	0.789	**0.648**	0.549	0.867	0.695	0.603
Ensemble on fixed split	0.826	0.632	0.561	**0.899**	0.696	0.614
Bagging Ensemble	**0.846**	0.637	**0.568**	0.898	**0.699**	**0.621**

Table 2

Detection results for the models trained on Kvasir-SEG dataset. Results in the last line corresponds when bagging ensemble approach trained on EndoCV2021 dataset and tested on Kvasir-SEG test set. Individual EfficientDet model results are given for their own cross-validation folds.

Model	Kvasir-SEG Test Set			EndoCV2021 as Test Set		
	mAP@50	mAP@75	mAP	mAP@50	mAP@75	mAP
EfficientDet-D0	0.860	0.696	0.616	0.563	0.411	0.369
EfficientDet-D1	0.856	0.685	0.586	0.602	0.434	0.383
EfficientDet-D2	0.882	0.689	0.603	0.577	0.422	0.371
EfficientDet-D3	0.883	**0.746**	0.634	0.614	0.468	0.411
Bagging Ensemble	0.904	0.737	**0.650**	**0.659**	**0.503**	**0.446**
Bagging Ensemble (EndoCV2021)	**0.906**	0.712	0.631	-	-	-

if there was no decrease in the validation set loss for the last 25 epochs. $2\times$ NVIDIA RTX 2080 GPUs are used for the training of D0 and D1 models and $4\times$ NVIDIA V100 16GB GPUs are used for the training of D2 and D3 models. Input size for each model and their inference speed, in terms of frames-per-second (FPS) on NVIDIA RTX 2080 GPU are given in Table 3. The processing time per frame is calculated as the sum of data transfer time from CPU to GPU, forward-pass processing through the model, and applying post-processing steps (confidence thresholding and NMS operation). Weighted box fusion adds 0.012 seconds processing time on average. Processing speeds were calculated by taking the average for 1000 images.

5. Experimental Results & Discussion

Experimental results in Table 1 show that the performances of the EfficientDet models increases in line with their scales, as expected. Fusion of different models, which are trained on a fixed training-validation split, through weighted box fusion results in increased performance. The best overall detection performance on both test sets is obtained when the proposed bagging ensemble approach is used. A similar trend is observed when the proposed approach is tested on the Kvasir-SEG dataset (Table 2). In addition, bagging ensemble architecture trained on

the EndoCV2021 dataset gets a very competitive accuracy on the Kvasir-SEG test set. When EfficientDet and bagging ensemble models, which are trained on the Kvasir-SEG dataset, are tested on the EndoCV2021 dataset, the bagging ensemble gets the best detection performance. This shows that, in addition to improving the performance on the same distribution test set, bagging ensemble approach also improves the detection performance on an out-of-distribution test set.

Experiment results in Table 1 and Table 2 show that, the bagging ensemble approach improves the detection performance on both datasets. In this approach, different bootstrap samples create perturbations for the unstable models and improve the ensemble performance in comparison to the fixed training set ensemble [23]. Moreover, in the case of fixed training-validation split for the ensemble, a significant part of the dataset is kept away for the validation set, limiting the number of samples that can be used for training. Bagging-style ensemble ensures that the fusion of different models trained on their respective training set covers the whole dataset available for training. We cross-dataset validated the proposed approach by training on EndoCV2021 and testing on Kvasir-SEG and vice versa. Results show that bagging-type ensemble improves the baseline results on out-of-distributions as well (mAP of 0.603 vs. mAP of 0.621 for training on EndoCV2021 and testing on Kvasir-SEG and mAP of 0.411 vs. mAP of 0.446 for training on Kvasir-SEG and testing on EndoCV2021). A significant observation in cross-dataset validation of out-of-distribution samples is that training on the EndoCV2021 dataset obtained better generalization performance compared to the training on Kvasir-SEG (mAP of 0.611 vs. mAP of 0.446). Two possible reasons for this result are training set size (1191 in EndoCV2021 vs. 800 in Kvasir-SEG) and dataset heterogeneity. Images in the EndoCV2021 dataset come from 5 different centers, which makes it a more diverse dataset in comparison to the Kvasir-SEG dataset, samples of which are from a single institution. This result highlights the importance of sample heterogeneity in the dataset for out-of-distribution performance. EndoCV2021 leaderboard also incorporated the generalisation metrics similar to detection generalisation defined in [30] to quantify the performance gaps.

An analysis of the detection results reveal that many glare artifacts are falsely detected as polyps (Figure 2 (b) and (c)). A separate artifact detector designed to work on endoscopic images [25, 31] may be incorporated into the system to reduce false positives and improve overall detection performance.
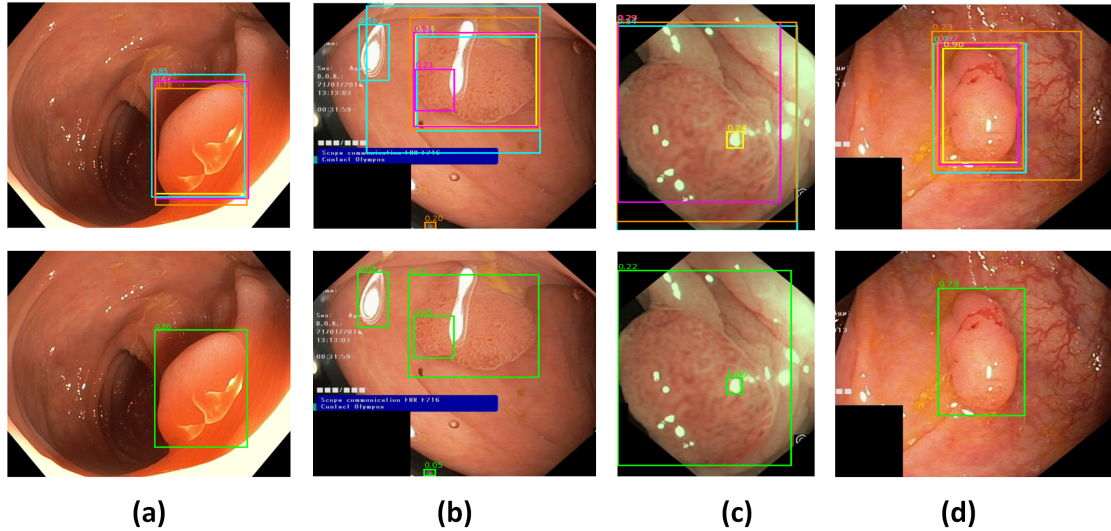
A drawback of the weighted box fusion is that it keeps all the bounding boxes on the image, similar to the affirmative type ensemble used in [25]. This type of ensemble tends to increase false positives; therefore, a consensus type ensemble, which only keeps the bounding boxes when the majority of the models agree, can be incorporated into weighted box fusion as future work.

With the advancements in hardware and software systems, parallel processing systems have become commonplace for real-time video processing applications [17]. If the individual models are run in parallel, the inference time will be upper bounded by the slowest model, which is EfficientDet-D3. Then, adding the weighted box fusion processing time on top of inference time, the total processing time is calculated as 65.2 milliseconds per frame, corresponding to 15.4 FPS, making it feasible for near real-time processing.

Table 3
Input sizes and FPS for different EfficientDet models.

Model	Input Size	FPS
EfficientDet-D0	512×512	40.6
EfficientDet-D1	640×640	29.9
EfficientDet-D2	768×768	26.9
EfficientDet-D3	896×896	18.8

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

Figure 2: Individual model predictions (first row) and their ensemble results (second row) on Kvasir-SEG dataset. Some models detects artifacts as polyps as can be seen in (b) and (c).

6. Conclusion

In this study, we have used a bagging-style ensemble method in association with different polyp detection networks. We have shown that using the proposed architecture increases the overall detection performance on both the same distribution test set and out-of-distribution test set, which is important for assessing the readiness of the methods for clinical use. While the proposed approach has been validated on one of the largest publicly available polyp datasets, it should be tested on larger and diverse datasets for a more comprehensive evaluation. The proposed method can be extended in the future by integration of an artefact detector and a consensus type ensemble may be used to reduce the number of false-positives. In addition, the use of higher scale EfficientDet models could be investigated, which could potentially provide better detection performance in exchange for higher computational cost.

Acknowledgments

References

[1] D. M. Parkin, Global cancer statistics in the year 2000, The Lancet Oncology 2 (2001) 533–543.

[2] WHO, Cancer, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/cancer.

[3] S. J. Winawer, A. G. Zauber, M. N. Ho, M. J. O'brien, L. S. Gottlieb, S. S. Sternberg, J. D. Waye, M. Schapiro, J. H. Bond, J. F. Panish, et al., Prevention of colorectal cancer by colonoscopic polypectomy, New England Journal of Medicine 329 (1993) 1977–1981.

[4] A. G. Zauber, S. J. Winawer, M. J. O'Brien, I. Lansdorp-Vogelaar, M. van Ballegooijen, B. F. Hankey, W. Shi, J. H. Bond, M. Schapiro, J. F. Panish, et al., Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths, N Engl J Med 366 (2012) 687–696.

[5] D. A. Lieberman, D. K. Rex, S. J. Winawer, F. M. Giardiello, D. A. Johnson, T. R. Levin, Guidelines for colonoscopy surveillance after screening and polypectomy: a consensus update by the us multi-society task force on colorectal cancer, Gastroenterology 143 (2012) 844–857.

[6] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, E. Butruk, Quality indicators for colonoscopy and the risk of interval cancer, New England Journal of Medicine 362 (2010) 1795–1803.

[7] Y. Iwahori, A. Hattori, Y. Adachi, M. K. Bhuyan, R. J. Woodham, K. Kasugai, Automatic detection of polyp using hessian filter and hog features, Procedia computer science 60 (2015) 730–739.

[8] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, E. K. Oermann, Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study, PLoS medicine 15 (2018) e1002683.

[9] S. Hwang, J. Oh, W. Tavanapong, J. Wong, P. C. De Groen, Polyp detection in colonoscopy video using elliptical shape feature, in: 2007 IEEE International Conference on Image Processing, volume 2, 2007, pp. II–465.

[10] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, M. Tzivras, Computer-aided tumor detection in endoscopic video using color wavelet features, IEEE transactions on information technology in biomedicine 7 (2003) 141–152.

[11] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, Y.-H. R. Tsai, Automated polyp detection in colon capsule endoscopy, IEEE transactions on medical imaging 33 (2014) 1488–1502.

[12] S. Ameling, S. Wirth, D. Paulus, G. Lacey, F. Vilarino, Texture-based polyp detection in colonoscopy, in: Bildverarbeitung für die Medizin 2009, Springer, 2009, pp. 346–350.

[13] N. Tajbakhsh, S. R. Gurudu, J. Liang, Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE, 2015, pp. 79–83.

[14] Automatic polyp detection challenge, 2015. URL: https://endovis.grand-challenge.org/.

[15] J. Bernal, N. Tajkbaksh, F. J. Sánchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, et al., Comparative validation of polyp detection

methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge, IEEE transactions on medical imaging 36 (2017) 1231–1249.

[16] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, I. Balasingham, Automatic colon polyp detection using region based deep cnn and post learning approaches, IEEE Access 6 (2018) 40950–40962.

[17] P. Wang, X. Xiao, J. R. G. Brown, T. M. Berzin, M. Tu, F. Xiong, X. Hu, P. Liu, Y. Song, D. Zhang, et al., Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy, Nature biomedical engineering 2 (2018) 741–748.

[18] K. Pogorelov, O. Ostroukhova, M. Jeppsson, H. Espeland, C. Griwodz, T. de Lange, D. Johansen, M. Riegler, P. Halvorsen, Deep learning and hand-crafted feature based approaches for polyp detection in medical videos, in: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2018, pp. 381–386.

[19] D. Jha, S. Ali, H. D. Johansen, D. D. Johansen, J. Rittscher, M. A. Riegler, P. Halvorsen, Real-time polyp detection, localisation and segmentation in colonoscopy using deep learning, arXiv preprint arXiv:2011.07631 (2020).

[20] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, H. D. Johansen, Kvasir-seg: A segmented polyp dataset, in: International Conference on Multimedia Modeling, Springer, 2020, pp. 451–462.

[21] M. Tan, R. Pang, Q. V. Le, Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.

[22] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[23] L. Breiman, Bagging predictors, Machine learning 24 (1996) 123–140.

[24] R. Solovyev, W. Wang, T. Gabruseva, Weighted boxes fusion: Ensembling boxes from different object detection models, Image and Vision Computing 107 (2021) 104117.

[25] G. Polat, D. Sen, A. Inci, A. Temizel, Endoscopic artefact detection with ensemble of deep neural networks and false positive elimination., in: Proc. International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2020) in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI2020), volume 2595, 2020, pp. 8–12.

[26] A. Neubeck, L. Van Gool, Efficient non-maximum suppression, in: 18th International Conference on Pattern Recognition (ICPR'06), volume 3, IEEE, 2006, pp. 850–855.

[27] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-nms–improving object detection with one line of code, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 5561–5569.

[28] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, M. A. Riegler, P. Halvorsen, C. Daul, J. Rittscher, O. E. Salem, D. Lamarque, T. de Lange, J. E. East, Polypgen: A multi-center polyp detection and segmentation dataset for generalisability assessment, arXiv (2021).

[29] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[30] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, S. Albarqouni, X. Wang, C. Wang, S. Watanabe, I. Oksuz, Q. Ning, S. Yang, M. A. Khan, X. W. Gao, S. Realdon, M. Loshchenov, J. A. Schnabel, J. E. East, G. Wagnieres, V. B. Loschenov, E. Grisan, C. Daul, W. Blondel, J. Rittscher, An objective comparison

of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific Reports 10 (2020) 2748. doi:`10.1038/s41598-020-59413-5`.

[31] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N. T. Nguyen, D. Q. Tran, L. D. Huynh, N. Boutry, S. Rezvy, H. Chen, Y. H. Choi, A. Subramanian, V. Balasubramanian, X. W. Gao, H. Hu, Y. Liao, D. Stoyanov, C. Daul, S. Realdon, R. Cannizzaro, D. Lamarque, T. Tran-Nguyen, A. Bailey, B. Braden, J. E. East, J. Rittscher, Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002. doi:`https://doi.org/10.1016/j.media.2021.102002`.