ENDOSCOPIC DETECTION AND SEGMENTATION OF GASTROENTEROLOGICAL DISEASES WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

Adrian Krenzer, Amar Hekalo, Frank Puppe

Department of Artificial Intelligence and Knowledge Systems, University of Würzburg, Germany

ABSTRACT

Previous endoscopic computer vision research focused mostly on the detection of a singular disease like, e.g. polyps. The endoscopic disease detection challenge (EDD2020) extends this classification task by providing data for different diseases in various organs. The EDD2020 includes two sub-tasks[1]: (1) Multi-class disease detection: localization of bounding boxes and class labels for the five disease classes: Polyp, Barret's Esophagus (BE), suspicious, High Grade Dysplasia (HGD) and cancer; (2) Region segmentation: boundary delineation of detected diseases. In this paper, we describe our approach by leveraging deep convolutional neural networks (CNNs). We highlight the comparison of two general state-of-the-art object detection approaches. The first one is Single Shot Detection (SSD), and the second one are two-step region proposal based CNNs. We, therefore, compare two different models: YOLOv3 (SSD) and Faster R-CNN with ResNet-101 backbone. For the second task, we leverage the state-of-the-art Cascade Mask R-CNN with various backbones and compare the results. In order to minimize generalization error, we apply data augmentation; finally, we use knowledge from the endoscopic domain to further refine our models during post-processing and compare the resulting performances.

1. INTRODUCTION

Endoscopic vision is a procedure which covers many different areas and organs of the human body, such as the bladder, the stomach or the colon, allowing gastroenterologists to potentially discover a wide array of diseases and abscesses, like polyps, cancer and Barrett's esophagus. Naturally, in order to assure detection of all diseases and to improve the workflow, application of real-time detection using Deep Learning is becoming more prevalent. There have been previous publications with good results on real-time detection of endoscopic polyps using Single Shot Detector [1] based CNNs [2] as well as an anchor free approach called AFP-Net [3]. Existing work usually focuses on one disease class, like polyp or cancer detection, mostly due to lack of annotated data. The Endoscopic Disease Detection Challenge 2020 [4] partially solves this issue by providing endoscopic images of three different organs, namely colon, esophagus and stomach, with five disease classes. Additionally they provide corresponding bounding boxes for object detection as well as polygonal masks for image segmentation. In this paper we apply and train state-of-the-art Deep Learning models for both tasks using various architectures and comparing their performance.

2. DATASETS AND DATA ANALYSIS

In order to choose and prepare the right deep CNN for the task, we start by analyzing the given training data in detail. The EDD2020 challenge [4] provides a training data set for multi-class disease detection, which contains 386 endoscopic images labeled with 684 bounding boxes and 502 segmentation masks. While analyzing the data, we recognize class imbalance. Therefore we counted the occurrences for each class throughout the dataset based on the bounding boxes. The dataset has more than 200 images with polyps and BE but less than 100 samples for the three remaining classes respectively. So, it might be challenging to learn the correct assessment of the classes HGD, suspicious and cancer. This unbalanced sample distribution is one difficulty of the dataset and is therefore considered while choosing our model and it's hyperparameters. The second difficulty we recognize is the variation in box sizes. We therefore calculated the area of all the boxes. Most of the boxes have nearly the same mean area while the variation of the areas differs enormously, especially for the polyp class, where the standard deviation is significantly larger than within other classes.

Finally, for the segmentation task, for every image there are given masks specifying which regions are of interest which is done separately for each class. While most of the images belong to a unique class, some of them have several masks with overlapping regions, which is especially apparent for the "suspicious" class. The latter is often only part of a region of an already existing class. Hence this is a multi-class multi-label segmentation task with independent classes. We randomly split the dataset into 90% training and 10% validation set, where the best model is chosen by minimum

[1] https://edd2020.grand-challenge.org

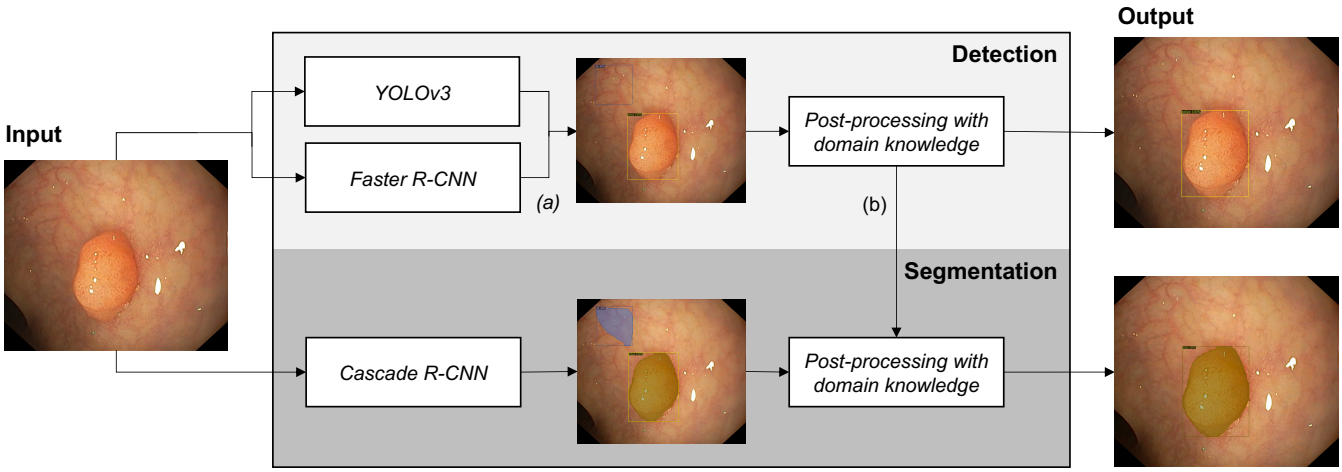

Fig. 1: This figure illustrates our final pipeline for the detection and segmentation task. At step (a) the predictions for polyps and HGD of the YOLOv3 algorithm and the predictions of BE, suspicious, and cancer of the Faster R-CNN are applied for the final result. At step (b) the box output of the detection architecture is utilized to filter the segmentation masks.

validation loss during training.

Additional data: In order to improve generalization, we extend the training dataset by including images from openly accessible databases. We include two datasets from a previous endoscopic vision challenge [5], namely the ETIS-Larib Polyp database [6], which consists of 196 polyp images, and the CVC-ClinicDB [7], which consists of 612 polyp images, as well as the dataset from the Gastrointestinal Image Analysis (GIANA) challenge [8], with 412 polyp images. All three datasets have corresponding segmentation masks. We add corresponding bounding boxes using the segmented masks ourselves. In addition we include the Kvasir-SEG dataset [9], which consists of 1000 polyp images with both segmentation masks and bounding boxes. Finally, we extract images annotated with esophagitis from the Kvasir2 dataset [10]. Esophagitis and Barret's esophagus occur at the same position in the esophagus, and some symptoms of esophagitis are very similar to Barret's esophagus symptoms. Therefore we add images with esophagitis symptoms which looked close to Barret's esophagus and test if those improve our results. We receive a light improvement in BE results and therefore include 103 additional images for a total of 2323 additional training images. Nevertheless, Barret's esophagus and esophagitis are different diseases and have to be distinguished in further research if more classes are included in the classification task.

3. METHODS

In this section, we illustrate our approaches for the two subtasks. All our models are trained on a Tesla P100 Nvidia GPU. After exploring the data, we decided to choose CNNs for the challenge as they have proven to be very stable in classic multi-class detection tasks like the COCO challenge [11]. In the domain of object detection, we consider two main concepts that have proven successful in multi-class object detection. First, a two-step method of region proposals and subsequent classification of the proposed regions like Faster R-CNN. Second single-shot detection (SSD), which is mostly applicable in real-time. We compare the results of the SSD model and Faster R-CNN. To improve our results further, we combine those two algorithms in our final architecture. For the second task, since both bounding boxes and segmentation masks are available, we choose the Cascade Mask R-CNN. Incorporating both types of annotations achieves the best results. For both of these tasks we add a post-processing with gastroenterological knowledge. Figure 1 depicts our final architecture for the detection and segmentation task. For training the Faster R-CNN we leverage the open source Detectron2 framework [12].

By including additional 2220 polyp images, we significantly increase the class imbalance of the training data. Class balance is crucial for training and inference of neural networks. To tackle this problem, we use class weights in the algorithms. Therefore the loss of an underrepresented class multiplies by a weight that balances the outcome of the total loss function. By adding those weights, we observe an enhancement in polyp detection while not losing the detection score in the other classes [13].

3.1. Task 1 multi-class bounding box detection:

As mentioned above, we want to compare two common object detection approaches, namely SSD and what we call a classic region proposal approach. Compared to classical approaches, SSD enables real-time detection. In practice, real-time detection is critical. Often, the gastroenterological diseases receive treatment directly (e.g., ablation of a polyp). Therefore

a low inference time has to be considered to apply the models in real practice. On the contrary, larger architectures may perform better in tasks suited for procedures like detecting the stadium of the disease, which mostly has no real-time restrictions. Nevertheless, a larger architecture may perform well on our challenge task, too. Therefore, we leverage one model from each of these sub positions. The model for SSD we utilize is called the YOLOv3 algorithm [14], which is the third version of the well-known YOLO architecture [15] and has added residual blocks that allow training deeper networks while preventing the vanishing gradient problem. We use the YOLOv3 algorithm with initial weights pre-trained on the COCO dataset [11]. In the next step, we unfreeze the last two layers of the network and train them utilizing the adam optimizer [16]. We train for 50 epochs. In addition, we unfreeze the whole network and train until it stops through early stopping, resulting in an additional 33 epochs.

As a classic larger architecture, we use a Faster R-CNN [17] with a 104 depth Retinanet backbone. We use a batch size of 2 because of the computational expense of this large network. We initialize the network with weights pre-trained on the COCO dataset. We choose a learning rate of 0.00025 for the training.

Post-processing: The YOLOv3 architecture is more successful in classifying polyps and HGD whereas classic architecture is better in detecting BE, suspicious and cancer. We therefore assemble both networks to improve our detection results. Hence, the YOLOv3 predicts HGD and polyps while the Faster R-CNN algorithm predicts BE, suspicious and cancer. Both algorithms can predict all labels, but we only use the predictions of the specified classes from each algorithm respectively. To further improve our results we use gastroenterological knowledge and knowledge of the data set structure. As the probability is low that BE and polyp are predicted in the same image we implement a simple rule: If both polyps and BE are detected, we only produce boxes for the class with higher probability, i.e., if the probability for polyps is higher than for BE, no bounding boxes are predicted for BE.

3.2. Task 2 region segmentation:

For the image segmentation task, we train two similar architectures with various backbones, namely Mask R-CNN [18] and its successor, Cascade Mask R-CNN [19]. Both architectures are primarily two-stage object detection models based on Faster R-CNN, i.e. a region proposal network first proposes candidate bounding boxes (Regions of Interest, RoI) before the final prediction. Here, they add another branch used to predict segmentation masks, where the proposed RoIs are used to enhance the segmentation mask predictions in contrast to using fully convolutional networks only. Cascade Mask R-CNN is an extended framework using a cascade-like structure and is essentially an ensemble of several Mask R-CNNs with weight sharing on the backbones.

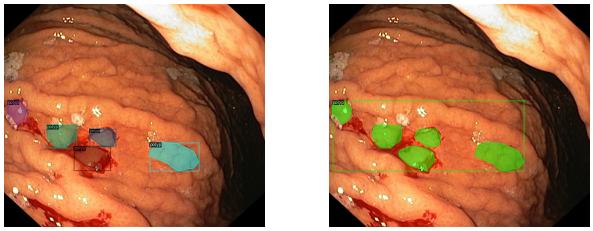

Fig. 2: In order to train Mask and Cascade Mask R-CNN for semantic segmentation, some bounding boxes had to be adjusted. We transform the boxes from including several instances (left) to be only one instance (right).

We choose these types of models for two reasons: First, since we have both bounding boxes and segmentation masks available as training data, we can utilize the Mask R-CNN approach, where RoI influences the segmentation, to the fullest. Second, since these networks are set to perform instance segmentation, each class is predicted independently from each other, which is a prefect fit for our multi-class multi-label problem. As this is a semantic task, we treat this as an instance segmentation with only one instance per occurrence per class. As such, we had to adjust some of the ground truth bounding boxes in our data, as shown in Fig. 2.

For Mask R-CNN we use the ResNeXt-101-32x8d [20] and for Cascade Mask R-CNN the ResNeXt-151-32x8d [20] models as backbones, both of which are CNN classifyers pre-trained on the ImageNet-1k dataset [21]. Additionally, both full architectures are pre-trained on the COCO dataset [11], hence we utilize transfer learning due to the small size of our training dataset.

The networks are trained using the Detectron2 framework [12] which provides a wide range of pre-trained object detection and segmentation models. As a pre-processing step, we convert our data to the COCO dataset format. Image pre-processing, i.e. padding, resizing, rescaling the pixel values etc., is then performed automatically within the framework. The total loss is the sum of classification, box-regression and mask loss $L = L_{\text{cls}} + L_{\text{box}} + L_{\text{mask}}$ [18], where L_{mask} is the binary cross-entropy for independent segmentation of all masks. The models are trained using stochastic gradient descent with a learning rate of 0.00025 and a batch size of 2. They are trained for up to 10000 iterations with checkpoints every 500 iterations. We then choose the checkpoint with the lowest validation loss as our final model. We also apply data augmentation in the form of random horizontal and vertical flipping as well as random resizing with retained aspect ratio in order to minimize the generalization error.

Post-processing: To further improve our results we use knowledge from gastroenterology and knowledge from the data set structure. As mentioned above, the probability that BE and polyps are present in the same image is very low. We apply the following procedure on the polyp/BE predictions:

- We utilize the predictions from object detection and only predict masks, where there are bounding boxes present from Yolov3 and Faster R-CNN.
- As an additional criterion, pixels within bounding boxes of probability < 0.2 are labeled with 0, i.e. no disease present.
- If both polyps and BE are detected, we only produce masks for the class with higher probability, as with the detection model.

4. RESULTS

In this section, we describe our results of the two subtasks. In both settings, we highlight the performance of the algorithms for every single disease. Therefore, we create a validation set. The validation set consists of 40 images randomly chosen from the provided data (no additional data is included). We test the detection as well as the segmentation on the created validation set.

4.1. Task 1

Table 1 shows our results on our created validation set for the detection task where YOLOv3 is the described SSD algorithm, Faster R-CNN is the FASTER R-CNN algorithm with ResNet-101 backbone and ensemble with pp (post-processing) is the ensemble of those two added with the hardcoded rule. We display the mean average precision with a minimum IoU of 0.5 (mAP) [11]. We highlight the performance of the algorithms split on the five diseases. All of the algorithms have an excellent performance in detecting polyps; this is mostly due to our additional polyp training data (see chapter 2). BE is better detected by the Faster R-CNN algorithm, which is why we used this algorithm for detecting BE in the ensembled version. Notably, suspicious is one of the harder classes to correctly classify as YOLOv3 is only showing a detection performance of 10 % mAP. As depicted in Table 1, cancer is detected quite well by all of the algorithms. All things considered, the ensemble with post-processing is the best algorithm in this task. The post-processing and combination of YOLOv3 and Faster R-CNN (Ensemble with pp) enhances the performance compared to the single YOLOv3 method by 7.95%. Figure 3 shows a detection result of the YOLOv3 algorithm and a segmentation result of the Cascade Mask R-CNN. Our detection score on the EDD2020 challenge [4] test set using the ensemble architecture produces a score of 0.3360 ± 0.0852.

4.2. Task 2

As in task 1, we evaluated our models on our validation set as a subset of the provided data on both Dice coefficient as well as intersection over union (IoU). Table 2 summarizes these

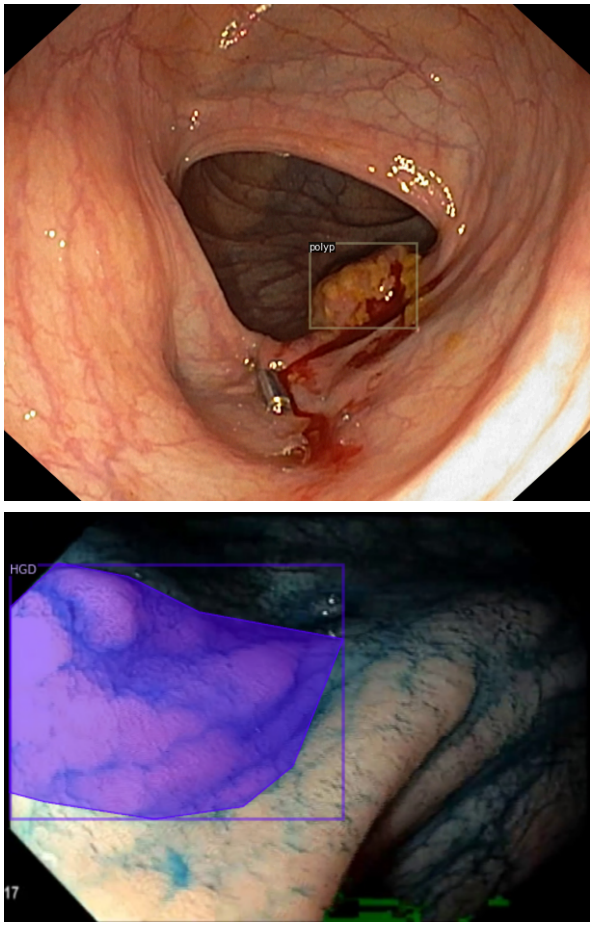

Fig. 3: Exemplary results for both detection with YOLOv3 (upper) and segmentation with Cascade Mask R-CNN (lower)

Table 1: Detection results on the validation data (mAP). MAP is the mean average precision over the five classes. Ensemble$_{pp}$ denotes the ensemble of YOLOv3 and Faster R-CNN with additional post-processing. All values are in %.

	YOLOv3	Faster R-CNN	Ensemble$_{pp}$
Polyp	84.19	73.50	84.46
BE	38.25	50.40	50.88
Suspic.	10.00	33.70	33.70
HGD	39.98	28.31	39.98
Cancer	49.99	53.20	53.20
mAP	44.49	37.29	**52.44**

results. While Mask R-CNN outperforms Cascade Mask R-CNN in both polyp and BE classes, Cascade Mask-RCNN provides better results overall, especially on the other three classes, which are comparatively underrepresented in our training data. Applying the post processing steps described in section 3 further improves the results of Cascade Mask R-CNN, but interestingly worsens the micro (μ) averaged score,

Table 2: Segmentation results on the validation data. R-CNN$_M$, R-CNN$_{CM}$ and R-CNN$_{CMpp}$ denote Mask R-CNN, Cascade Mask R-CNN and Cascade Mask R-CNN with post processing respectively. We also computed the micro averaged scores, denoted by μ mean, in contrast to mean, which is averaged over class scores. All values are in %.

	R-CNN$_M$		R-CNN$_{CM}$		R-CNN$_{CMpp}$	
	Dice	IoU	Dice	IoU	Dice	IoU
Polyp	69.41	67.03	61.57	60.08	69.07	67.58
BE	46.41	43.84	44.48	41.06	46.56	43.08
Suspic.	27.64	25.94	40.03	38.83	52.53	51.33
HGD	41.83	38.28	63.59	60.25	68.25	65.75
Cancer	53.77	52.14	55.86	54.96	57.24	57.00
mean	47.81	45.45	53.11	51.04	**58.73**	**56.95**
μ mean	36.57	27.05	**47.66**	**38.44**	45.36	37.17

which we discuss below. Our segmentation score on the EDD2020 challenge [4] test set using Cascade Mask R-CNN is then 0.6526 ± 0.3418.

5. DISCUSSION & CONCLUSION

All of our models in both tasks perform best on the polyp class and worst on the suspicious category. Since data on polyps is abundant in our training set, it is clear why the networks show good results in this area. The suspicious class, however, shows a similar amount of samples as HGD and cancer, yet, with the exception of Cascade Mask R-CNN, all models perform significantly worse on this class. This is most likely due to the unclear nature of this class as it often denotes regions belonging to different types of diseases, i.e. in some images it denotes possible cancer, whereas in others it signifies possible BE. Additionally, performing gastroenterologists often have differing opinions on what areas can be considered as suspicious, which adds further noise to our data. The performance of Cascade Mask R-CNN on suspicious and the other less represented classes can be attributed to its ensemble-like structure. The discrepancy of the micro-averaged scores can be explained as such: Our post processing severely reduces the amount of false positives, but also adds some false negatives. This improves the class-based score, since classes on one image with empty masks receive perfect scores this way. With micro-averaging, however, since precision and recall are the same, we essentially look at the per pixel accuracy of the entire mask, ultimately worsening this score.

Our model outperforms the best network from [2], namely SSD with a InceptionV3 backbone, which was partially trained using the same polyp databases and showed a precision of 73.6% on the MICCAI 2015 evaluation dataset, compared to our 84.19% with YOLOv3. AFP-net performs better than our model [3] with a precision of 88.89% on the ETIS-Larib dataset and 99.36% on the CVC-Clinic-train dataset. However, for both cases, direct comparison is difficult since both different training and different evaluation data are used. Additionally, we perform multi-class prediction, which can be a more difficult task to perform than binary prediction.

We applied state-of-the-art Deep Learning architectures for the detection and semantic segmentation of five different gastroenterological diseases. For detection, we evaluated three architectures, the YOLOv3 and the Faster R-CNN, and our combination of those algorithms. Furthermore, our ensemble includes domain knowledge-based post-processing, which further enhances our results in the challenge. For segmentation, we evaluate three models: Cascade Mask R-CNN, its predecessor Mask R-CNN, and the Cascade Mask R-CNN combined with post-processing. In the region segmentation task, the Cascade Mask R-CNN with additional post-processing reliably performs as good or better than the other networks. For future work we intend to improve our results by adding more training data, applying additional forms of data augmentation and further hyperparameter tuning. All in all, we present state-of-the-art results in the EDD challenge with our detection and segmentation applications.

6. REFERENCES

[1] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. *CoRR*, abs/1512.02325, 2015.

[2] J. Jiang M. Liu and Z. Wang. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. *IEEE Access*, 7:75058–75066, 2019.

[3] Dechun Wang, Ning Zhang, Xinzi Sun, Pengfei Zhang, Chenxi Zhang, Yu Cao, and Benyuan Liu. Afp-net: Realtime anchor-free polyp detection in colonoscopy, 2019.

[4] Sharib Ali, Noha Ghatwary, Barbara Braden, Dominique Lamarque, Adam Bailey, Stefano Realdon, Renato Cannizzaro, Jens Rittscher, Christian Daul, and James East. Endoscopy disease detection challenge 2020. *arXiv preprint arXiv:2003.03376*, 2020.

[5] J. Bernal, N. Tajkbaksh, F. J. Snchez, B. J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. Maier-Hein, S. Speidel, D. Stoyanov, P. Brandao, H. Crdova, C. Snchez-Montes, S. R. Gurudu, G. Fernndez-Esparrach, X. Dray, J. Liang, and A. Histace. Comparative validation of polyp detection methods in video colonoscopy: Results from the miccai 2015 endoscopic vision challenge. *IEEE Transactions on Medical Imaging*, 36(6):1231–1249, June 2017.

[6] J. Silva, A. Histace, O. Romain, et al. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *Int J CARS*, 9:283 – 293, 2014.

[7] 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.

[8] Y. B. Guo and Bogdan J. Matuszewski. Giana polyp segmentation with fully convolutional dilation neural networks. In *VISIGRAPP*, 2019.

[9] Debesh Jha, Pia H. Smedsrud, Michael Riegler, Pål Halvorsen, Dag Johansen, Thomas de Lange, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*. Springer, 2020.

[10] Konstantin Pogorelov, Kristin Ranheim Randel, Carsten Griwodz, Sigrun Losada Eskeland, Thomas de Lange, Dag Johansen, Concetto Spampinato, Duc-Tien Dang-Nguyen, Mathias Lux, Peter Thelin Schmidt, Michael Riegler, and Pål Halvorsen. Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, pages 164–169, New York, NY, USA, 2017. ACM.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[13] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016.

[14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[19] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *CoRR*, abs/1906.09756, 2019.

[20] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. *CoRR*, abs/1611.05431, 2016.

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.