ARTEFACT DETECTION AND SEGMENTATION USING CASCADE R-CNN & U-NET

Hoang Manh Hung, Phan Tran Dac Thinh*, Hyung-Jeong Yang, Soo-Hyung Kim, and Guee-Sang Lee*

Department of Electronic and Computer Engineering, Chonnam National University, South Korea

ABSTRACT

Endoscopy is a widely adopted procedure for the early detection of various types of cancers, therapeutic procedures and minimally invasive surgery. Nevertheless, the efficiency of this method is badly influenced by several artefacts, namely pixel saturations, motion blur, defocus, bubbles, specular reflections, etc. Providing that all of these artefacts are spotted comprehensively, contaminated frame could be restored to an adequate quality in order to better visualize the underlying tissue during diagnosis. In this paper, we present and discuss our methodology of detection and segmentation on endoscopy images. For artefact detection, we modified a deep neural network structure based on the cascade R-CNN with ResNeXt101 as backbone, including the deformable convolutions. Moreover, Feature Pyramid Network is added to refine the raw feature maps, enhancing the performance of feature extraction. In semantic segmentation task, U-Net is utilized with the support of SE-ResNeXt50 as backbone. The classification model appears to be legitimately dominant when compared to other models that we tested. At the end of the Endoscopy Artefact Detection challenge 2020, we attain the mAP score of 0.2366 with the deviation of 0.0762 on the test dataset of detection task and the dice score of 0.5700 on the test dataset of segmentation.

1. INTRODUCTION

In the last decade, medical imaging for disease diagnosis and early treatment has a great step forward thanks to the surge of machine learning application in computer vision. The applications of this new technique are indisputably innumerable, shortening the time of diagnose and accelerating the treatment. Although the precision of the software related to medical imaging currently still cannot be compared to that of experts, recent advances have been proving its auspicious capabilities of human replacement in differing tasks. Endoscopy is one of those clinical procedure that can utterly benefit from this burgeoning technology. In this procedure, a long, thin, flexible tube along with a light source and a camera at the tip captures the frames inside your body, observing the uncommon specks which could be indications of disease or cancer. Unfortunately, diagnosis is made difficult by corrupted frames with multiple artefacts such as pixel saturations, motion blur, defocus, fluid, debris, etc. However, a technique called high-quality endoscopic frame restoration is able to thoroughly solve this issue if undesired objects are accurately tracked down. As a result, the task of detecting and identifying those artefacts is crucial for the procedure of endoscopic diagnosis. An amendment technique called high-quality endoscopic frame restoration is able to thoroughly solve this issue only if undesired objects are accurately tracked down. Consequently, the task of detecting and identifying those artefacts is crucial for the procedure of endoscopic diagnosis. Each endoscopic frame is tainted by multiple artefacts and the influence of them on each image is not even. Unless the restoration technique has the knowledge of those artefacts precise spatial location, the quality of image can be guaranteed for further diagnosis.

Object detection and segmentation has been gaining a lot of attraction lately, leading to the advent of many powerful neural network models. In terms of object detection, Zhang et al proposed Mask-Aided R-CNN based on Mask R-CNN [1] to modify its mask header for assisting training on pixel-level labelled samples. RetinaNet [2] associated with focal loss was introduced by Lin et al in order to predict bounding boxs sizes more precisely and cope with class imbalance issue. To reduce vanishing positive samples for large thresholds and increase hypotheses quality, Cascade R-CNN [3] with feature pyramid networks (FPN) [4] was presented by Cai et al. For object segmentation, U-Net [5] built by Ronneberger et al is one of the most popular models. It is favored for multiple of applications and performs effectively. Moreover, DeepLab [6] proposed by Chen et al is a vigorous model because it is capable of enlarging the field of view of filters for larger context learning but not compromise with the amount of computation. With its high capability of global context information exploitation, Pyramid Scene Parsing (PSP) [7] network introduced by Zhao et al is another model that achieved high record in the task of segmentation.

Endoscopy Artefact Detection 2020 (EAD) Challenge [8, 9, 10] is one of the competitions that are interested in optimizing the automatic artefact detection capability. The challenge contributes two kinds of labelled data: for detection task, they

*Both authors contributed equally to this manuscript

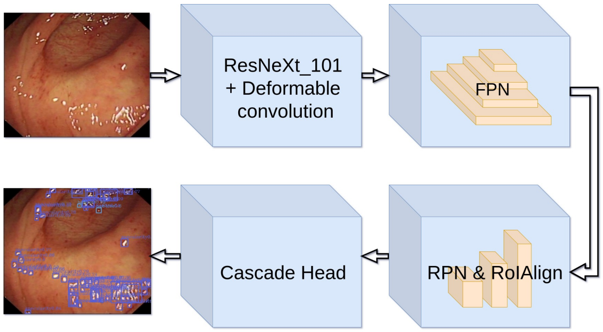

Fig. 1. The proposed method based on Cascade R-CNN.

provide images with bounding box annotation and for segmentation task, the data are images with pixel-level annotation. In this paper, we present two different models for each task of the EAD challenge after several preliminary tests with a few models. Cascade R-CNN is utilized for former task and U-Net with SE-ResNext50 is applied for the latter task. Both models are discussed in Section 2 and we delineate the progress of training them in Section 3.

2. METHODS

2.1. Artefact Detection

Our approach is based on Cascade R-CNN [3], which is trained sequentially by using cascaded bounding box regression. This network can produce higher quality proposals during the inference process than other models that we have tested. In addition, the detector uses the resampling mechanism which can reduce overfitting with the high intersection over union (IoU) threshold and also dismiss some outliers. The backbone ResNeXt-101 (64x4d), following by the Feature Pyramid Networks (FPN) [4] in a top-down mode as the neck, increases the networks capability of feature extraction and improves recall. Furthermore, the deformable convolution (DCN) [11] is added into the backbone stage 3 to stage 5, which assists the model in perceiving the image content and differentiating the desired objects from the large background. Our modified model for detection is illustrated in Fig. 1.

2.2. Semantic Segmentation

The mask images for segmentation contain 5 classes, namely instrument, specularity, artefact, bubbles and saturation. The distribution of desired objects in each images is not even and some artefacts doesnt appear much on the background, especially the case of specularity. Moreover, sometimes, the mask images are ambiguous when we try to point out the foreground and background. Therefore, we need a tool that is powerful enough to solve the previously-mentioned problems. In this task, after assessing several models for segmentation, we decide to take the advantage of segmentation from U-Net [5] that have been verified through a lot of papers which include papers about medical segmentation. Many researches favored U-Net as their main network. Usually, U-Net is preferred mostly because of its flexible and interchangeable backbone. The competitive classification performance of the neural network as backbone will determine the success of the whole segmentation model. The higher the accuracy of the backbone can achieve, the better the U-Net model can tell the difference between desired pixels and background. Therefore, through a few trials with some backbones (variants of Resnet, ResNext and SE-ResNeXt), we choose SE-ResNeXt50 model because of its equilibrium among strong performance, reasonable computational cost and acceptable time of training. Some models yielded greater results than our chosen model but they took us a huge amount of time to train. SE-ResNeXt50 is a modified version of ResNeXt model with Squeeze and Excitation blocks [12], improving the representational power of a network. This model can tackle the problem of the imbalanced dataset better than other models and retrieve more hidden fragments inside the picture. Furthermore, binary cross entropy loss and dice loss are combined to deal with the discrete distribution of the foreground. Another important point is using the pretrained model of the backbone. Not only does it increase the accuracy of prediction but also reduces the total amount of training time. The predicted mask is subsequently applied a threshold value which is verified earlier. This threshold value from 0.2 to 0.9 is picked if it satisfies the best result of dice score on the total dataset.

3. EXPERIMENTAL RESULTS

3.1. Artefact Detection

In multi-class artefact detection task, the dataset consists of 2531 images and each image can contain one or many classes in the total of 8 classes, namely specularity, saturation, artifact, blur, contrast, bubbles, instrument and blood. First, we preprocess the images by transforming all of them in to the size of 608x608. Then, they are normalized by our mean and standard and augmented with random flip and crop. Stochastic gradient descent is the chosen optimization along with the initiated learning rate of 0.02. The backbone uses the pretrained model on COCO dataset (2017) to enhance the performance of the main model. The 5-fold cross validation is used for stable deviation. In the testing phase, weights from training 5 folds are used to detect the online testing dataset. Non-max suppression (NMS) [13] is selected to filter redundant detections in post-processing for each fold. Subsequently, to combine the predictions from 5 weights, we employ Weighted Boxes Fusion (WBF) [14] method instead of the commonly NMS. WBF not only deletes redundant boxes but also takes

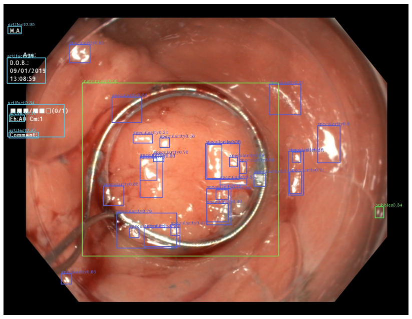

Fig. 2. Sample of predicted image from test dataset.

Table 1. Detection results.

Dataset	Model	Score_d	gmAP
Test	Cascade R-CNN + DCN + ResNeXt101	0.2366	0.2153

advantage of information from these boxes to align the final boxes more accurately. At the end, we get mAP score of 0.2366 with a deviation of 0.0762 in the testing set as shown in Table 1. The Fig. 2 below is one of our predicted images from the test dataset.

3.2. Semantic Segmentation

The challenge provides 643 pixel-level annotation images for semantic segmentation. The size of images is not fixed so it needs to be preprocessed before training. There is no separate validation data so we use 5-fold cross validation to ensure the correctness of our own evaluation. First, we set the size of all images into 256x256. The original size of images is varied and bigger than the normalized size. Empirically, if the normalized size is set at higher value, the overall result of segmentation task is absolutely enhanced. Second, we train the U-Net segmentation model with the backbone network of SE-ResNeXt50. We also use the pretrained model of SE-ResNext50 to boost the performance of segmentation and reduce the duration of training. The backbone is not frozen for fine-tuning at the beginning. We load the weight of the backbone and train it at the same time with the main model. Augmentation is not applied in our training method due to lower results when it is compared to training without augmentation on a few lighter-weight neural network models. We did not check if augmentation could boost the performance of the model due to lack of time. The GPU we have is Nvidia RTX 2080Ti so we set the batch size of 32. Adam optimizer is used for our network and the learning is kept constant at 3x10^{-4}. Due to overlay in pixel-level labels, which means one pixel could contain more than one class, we train each class once at the same time. Thus, we have to train 25 times in total for 5 classes, which is evaluated by 5 fold cross validation. The amount of epochs for one training time is 100. The predicted mask of one class is the average score of 5 model weights from 5 folds. After training, we find the threshold for our predicted images. 0.4 is the threshold value which leads us to our own best score as shown in Table 2.

Table 2. Segmentation results.

Dataset	Model	sscore	sstd
Test	U-Net + SE ResNeXt50	0.5700	0.2703

4. CONCLUSION

In this paper, we demonstrate two verified models for the EAD2020 Challenge. The difficulty that we met when participating in this challenge is how to pick up the most suitable component for our network that performs excellently on the given dataset. It proves that there are still a lot of approaches that we havent tested and the development of artefact detection related to endoscopy is very promising.

5. ACKNOWLEDGMENT

This research was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) & funded by the Korean government (MSIT) (NRF-2019M3E5D1A02067961) and a grant (HCRI 19136) Chonnam National University Hwasun Hospital Institute for Biomedical Science and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2017R1A2B4011409).

6. REFERENCES

[1] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42, 2020.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollr. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[4] Tsung-Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, Bharath Hariharan, , and Serge Belongie. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science: Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351, 2015.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 2018.

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. Endoscopy artifact detection (ead 2019) challenge dataset. *arXiv preprint arXiv:1905.03209*, 2019.

[9] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James East, Xin Lu, and Jens Rittscher. A deep learning framework for quality assessment and restoration in video endoscopy. *arXiv preprint arXiv:1904.07073*, 2019.

[10] Sharib Ali, Felix Zhou, Barbara Braden, Adam Bailey, Suhui Yang, Guanju Cheng, Pengyi Zhang, Xiaoqiong Li, Maxime Kayser, Roger D. Soberanis-Mukul, Shadi Albarqouni, Xiaokang Wang, Chunqing Wang, Seiryo Watanabe, Ilkay Oksuz, Qingtian Ning, Shufan Yang, Mohammad Azam Khan, Xiaohong W. Gao, Stefano Realdon, Maxim Loshchenov, Julia A. Schnabel, James E. East, Geroges Wagnieres, Victor B. Loschenov, Enrico Grisan, Christian Daul, Walter Blondel, and Jens Rittscher. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports*, 10, 2020.

[11] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[13] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. *18th International Conference on Pattern Recognition (ICPR)*, 3, 2006.

[14] Roman Solovyev and Weimin Wang. Weighted boxes fusion: ensembling boxes for object detection models. *CoRR*, abs/1910.13302, 2019.