ENDOSCOPIC ARTEFACT DETECTION USING CASCADE R-CNN BASED MODEL

Zhimiao Yu, and Yuanfan Guo

Shanghai Jiao Tong University, Shanghai, China
{gyfastas,Carboxy}@sjtu.edu.cn

ABSTRACT

Accurate detection of artefacts is a core challenge in a wide-range of endoscopic applications addressing multiple different disease areas. Our work aims to localise bounding bboxes and predict class labels of 8 different artefact classes for given frames and clinical endoscopy video clips. To solve the task, we use Cascade R -CNN[1] as network architecture and adopt ImageNet pretrained ResNet101[2] as backbone with Feature Pyramid Network (FPN) [3] structure. To improve the network performance, methods like data augmentation and multi-scale are also be adopted. In the end, we analyze the major challenge of the task.

1. INTRODUCTION

Endoscopy is a widely used clinical procedure for the early detection of numerous cancers (e.g., nasopharyngeal, oesophageal adenocarcinoma, gastric, colorectal cancers, bladder cancer etc.), therapeutic procedures and minimally invasive surgery (e.g.,laparoscopy). However, video frames captured by an endoscope usually contain multiple artefacts, which not only present difficulty in visualising the underlying tissue during diagnosis but also affect any post analysis methods required for follow-ups. Existing endoscopy workflows are not competent qualified for restoring high-quality endoscopic frames because they can detect only one artefact class in most cases. Generally, the same video frame can be corrupted with multiple artefacts, e.g. motion blur, specular reflections, and low contrast can be present in the same frame. Besides, corruption varies with video frames in artefact types. Therefore, improving detection accuracy is a core challenge in a wide-range of endoscopic applications.

Recently, deep ConvNets have significant improved image classification and object detection accuracy[4]. In deep learning era, object detection can be grouped into two genres: “two-stage detection” (e.g. RCNN[5]) and “one-stage detection” (e.g. [6])[7]. In this task, we use Cascade R-CNN[1] as network architecture. It is a multi-stage object detection architecture. The reason we adopt Cascade R-CNN as our network architecture is it achieves state-of-art detection performance.

2. DATASETS

The 8 artefact classes in the dataset for “Endoscopic Artefact Detection” include specularity, specularity saturation, artifact, blur, contrast, bubbles, instrument and blood. The visualization of ground truth bboxes are shown in Fig 2. The artefact detection task will be evaluated based on the results of the test dataset provided from a subset of the data collected for training. Specifically, the training dataset for detection consists in total 2200 annotated frames over all 8 artifact classes and test dataset 100[8] [**?**] [9].

3. METHODS

3.1. Architecture

The model architecture is shown in Fig 1. We use Cascade R-CNN[1] as network architecture and adopt ImageNet pretrained ResNet101[2] as backbone with Feature Pyramid Network (FPN)[3] structure. Taking the areas of artefacts into consideration, the anchors base areas are tuned from 16^2 to 512^2 on $P2$ to $P6$. Specifically, anchor scales, ratios and strides are [8], [0.5, 1.0, 2.0] and [4, 8, 16, 32, 64], respectively.

3.2. Implement Details

For data augmentation, each image will be horizontally flipped with a 50 percent chance. We replace the nms operation with the $soft$-nms[10] operation in the architecture and set the learning rate scheduling strategy as consine decay[11]. The classification and regression loss function are CrossEntropyLoss and SmoothL1Loss, respectively. The model is trained for 24 epochs.

In the experiment, we find that *specularity*, *artifact* and *bubbles* are hard to classify. A probable reason is these three artefacts have similar appearance (e.g. Some of them all appears as spots of light). To solve this problem, we modify the loss function. In specific, we up-weight loss when model mistakenly classify these three artefacts. The result turns out

---

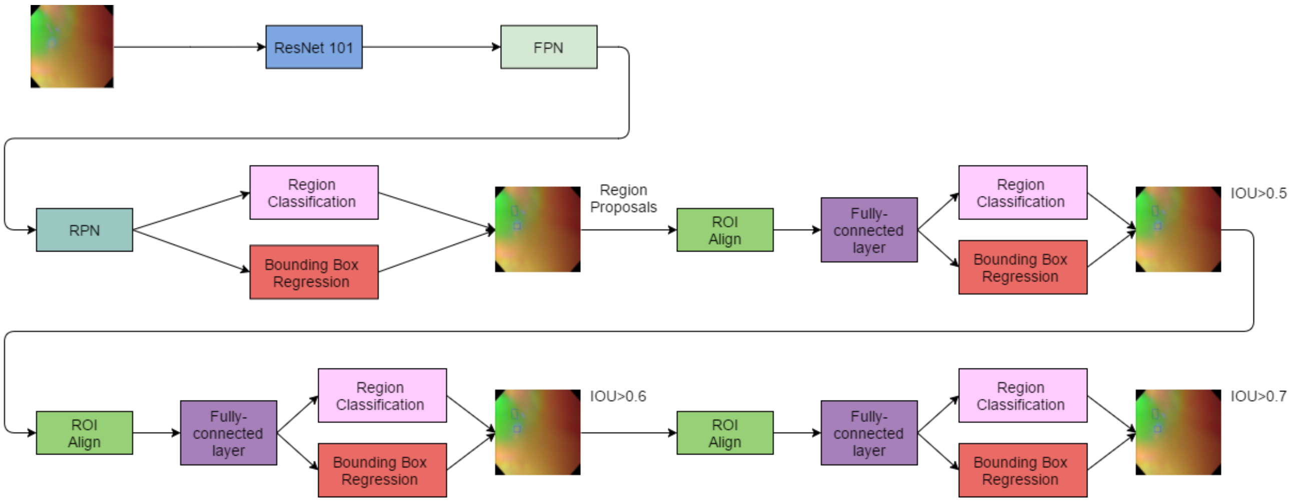

Fig. 1. Model architecture based on Cascade R-CNN.

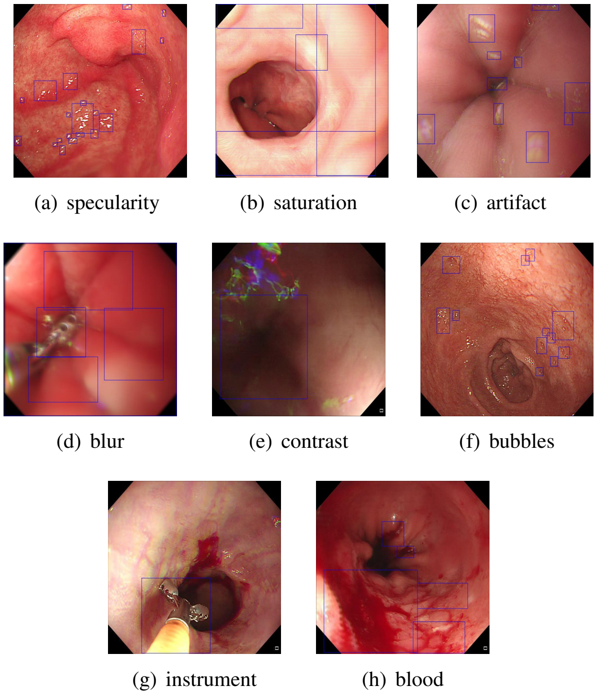
(a) specularity (b) saturation (c) artifact
(d) blur (e) contrast (f) bubbles
(g) instrument (h) blood

Fig. 2. Visualization of ground truth boxes.

to be an improvement of AP for these three artefacts but a decline of mAP.

4. RESULTS

We randomly divide the data provided into 5 subsets and use one of them for validation while others for training. The following metrics are based on the validation set.

4.1. Data augmentation of resizing

We report our results obtained from **baseline** in Table 1. Our baseline achieves 0.26 mAP. To improve the model performance, we added image resizing operation to the data augmentation pipeline. Specifically, each image will be randomly resized among the range from (512, 512) to (1024, 1024) with the same aspect ratio as the original. Considering the image size varies, we believe this operation will be effective.

The results are shown in Table 2. According to the results, we argue that resizing operation can obvious improve the model performance with an increase in mAP of 0.017. We notice that the improvement is mainly on AP^{small}. The main reason is that in most cases the resizing operation enlarges image size and thus makes it possible to detect more small objects.

Note that the scales of test images are often larger than train images, i.e. images in testset often with height and width larger than 1000 while images in trainset around 500, so the resizing operation can solve the scale mismatch problem between training images and testing images.

4.2. Difficult classification among specularity, artifact and bubbles

In the experiment, we found that network has some difficulties in distinguishing classes among *specularity*, *artifact* and *bubbles*. To demonstrate the problem clearly, we calculated the confusion matrix, which is shown in Table 4. According to the Table 4, the network has two drawbacks. Firstly, the network tends to confuse *specularity*, *artifact* and *bubbles* in the classification procedure. Secondly, the network has poor performance in detecting *blur*.

To solve the first problem, we modified the loss function. Specifically, we increased the loss weights to the misclassification of *specularity*, *artifact* and *bubbles*. The result

Table 1. Baseline performance on validation set.

Artefacts	AP	$AP^{IoU=.50}$	$AP^{IoU=.75}$	AP^{small}	AP^{medium}	AP^{large}
specularity	0.123	0.319	0.063	0.064	0.193	0.202
saturation	0.197	0.670	0.217	0.040	0.210	0.345
artifact	0.225	0.486	0.170	0.129	0.218	0.421
blur	0.184	0.275	0.167	0	0	0.191
contrast	0.414	0.760	0.416	0.033	0.187	0.439
bubbles	0.124	0.345	0.061	0.094	0.128	0.216
instrument	0.531	0.801	0.624	/	0	0.551
blood	0.181	0.454	0.103	/	0.079	0.221
mean	0.260	0.514	0.228	0.060	0.127	0.323

Table 2. Model performance on validation set with resizing operation.

Artefacts	AP	$AP^{IoU=.50}$	$AP^{IoU=.75}$	AP^{small}	AP^{medium}	AP^{large}
specularity	0.138	0.380	0.062	0.091	0.216	0.199
saturation	0.295	0.669	0.246	0.050	0.212	0.338
artifact	0.243	0.516	0.185	0.140	0.239	0.427
blur	0.181	0.279	0.178	0	0	0.188
contrast	0.422	0.760	0.424	0	0.224	0.443
bubbles	0.153	0.384	0.085	0.125	0.151	0.254
instrument	0.569	0.830	0.649	/	0.044	0.587
blood	0.212	0.495	0.172	/	0.130	0.244
mean	0.277	0.539	0.250	0.068	0.152	0.335

Table 3. Final result on leaderboard.

dataset	dscore
50% testset	0.2603
100% testset	0.2036

turned out to be an improvement of AP for these three artefacts but a decline for mAP.

4.3. Qualitative Results

To find out what kinds of artefact our model can successfully detect, we show some qualitative results in Fig 3 and Fig 4. The qualitative results indicate a). for artefacts with not so small size, our model tends to generate accurate detections; b). more artefacts in an image lead to more difficulties in detecting; c). our model generates a fair number of true negative *blur*. We are not sure the reason for problem c) mentioned above is whether the shortcomings of the model itself or the absence of annotation *blur*, because the corresponding images show blur characters.

4.4. Leaderboard Result

We added image resizing operation to the data augmentation pipeline and fine-tuned the maximum box number per image to 300. Then we used the model to obtain the testset results and the performance is shown in Table 3.

5. DISCUSSION & CONCLUSION

In our work, we found the major challenge in “Endoscopic Artefact Detection” task is the difficult classification among *specularity*, *artifact* and *bubbles*. One intuitive explanation is that some of them all appears as spots of light, sharing a high degree of similarity. In the future, we intend to train 3 separate classifiers for these 3 artefacts and adopt more advanced feature extraction networks, which may solve this challenge to some extent. Boxes ensemble method was performed in our experiment. However, it seemed this method caused lower mAP.

To sum up, we constructed a Cascade R-CNN based model to solve the “Endoscopic Artefact Detection” task. We adopted several methods to improve the network performance, including data augmentation, modifying loss function and boxes ensemble. We also identified the major challenge in this task.

6. REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceed-*

Table 4. Confusion matrix of 8 classes

		Labels							
		specularity	saturation	artifact	blur	contrast	bubbles	instrument	blood
Predicted	specularity	**70.3%**	7.2%	**10.9%**	0.0%	0.0%	**19.9%**	0.0%	0.0%
	saturation	1.4%	79.9%	0.9%	0.7%	0.0%	0.5%	1.7%	0.0%
	artifact	**19.4%**	5.1%	**75.4%**	**20.6%**	1.9%	**13.1%**	6.0%	2.3%
	blur	0.0%	1.3%	2.3%	**49.1%**	3.8%	0.1%	4.3%	2.3%
	contrast	0.0%	0.0%	1.0%	**11.8%**	88.5%	0.0%	11.2%	12.8%
	bubbles	**8.4%**	2.6%	**8.7%**	0.0%	0.0%	**66.2%**	0.0%	0.8%
	instrument	0.3%	3.3%	0.7%	12.5%	4.5%	0.1%	75.0%	3.8%
	blood	0.2%	0.5%	0.1%	5.2%	1.2%	0.2%	1.7%	78.2%

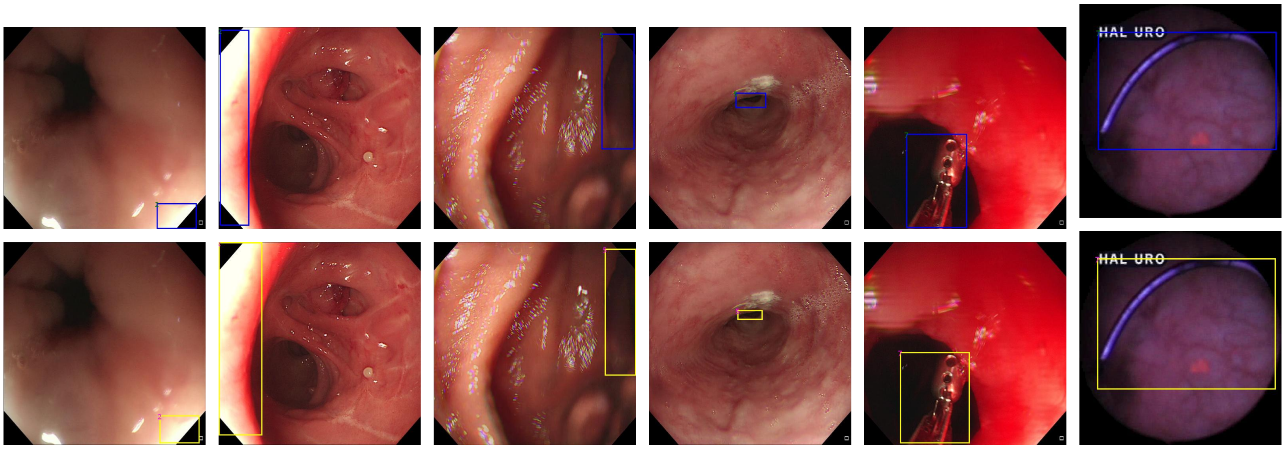

Fig. 3. High quality detection examples (i.e. Model generates accurate detections). The first row shows ground truth where artefacts are annotated with blue bounding boxes . The second row shows results where detected artefacts are annotated with yellow bounding boxes.

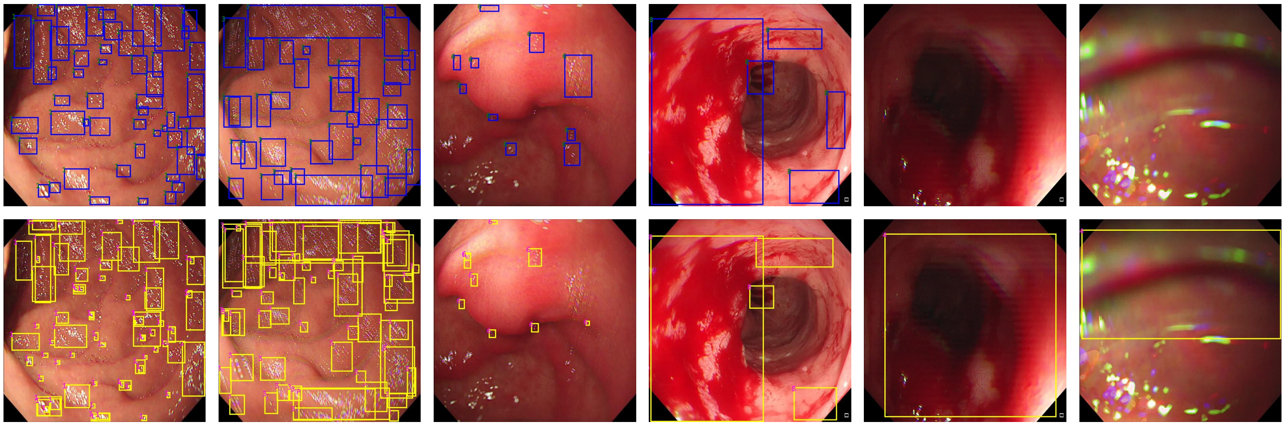

Fig. 4. Low quality detection examples (i.e. Model generates inaccurate detections). The first row shows ground truth where artefacts are annotated with blue bounding boxes . The second row shows results where detected artefacts are annotated with yellow bounding boxes. The last two columns represent false positive *blur*.

ings of the IEEE conference on computer vision and pattern recognition, pages 6154–6162, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision*

and pattern recognition, pages 770–778, 2016.

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):142–158, 2015.

[6] 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.

[7] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.

[8] 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.

[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] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[11] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 558–567, 2019.