**Department of Electrical Engineering**

**Indian Institute of Technology Roorkee**

**Industry Oriented Project (EEN-300)**

**A Technical Report on**

# Endoscopy Artefact Detection

**Name of Students:** 1. Name: Harshit Khetan

Enrolment No: 17115038

2. Name: Parv Bhatt

Enrolment No: 17115063

3. Name: Vibhore Mendiratta

Enrolment No: 17115110

**Project Supervisor:**

Dr. G N Pillai

Head of Department

Department of Electrical Engineering

Indian Institute of Technology Roorkee

# Table of Contents

## ABSTRACT

Endoscopy is a widely used clinical procedure for detecting numerous cancers, therapeutic procedures & minimally invasive surgery. A significant drawback of these video frames is that they are heavily corrupted with multiple artefacts (e.g., pixel saturation, motion blur, defocus, specular reflections, bubbles, fluid, debris, etc.). These artefacts not only present difficulty in visualizing the underlying tissue during diagnosis but also affect any post analysis methods required for follow-ups. Accurate detection of artefacts is a core challenge in a wide range of endoscopic applications addressing multiple different disease areas. The importance of precise detection of these artefacts is essential for high-quality endoscopic frame restoration & crucial for realizing reliable computer-assisted endoscopy tools for improved patient care.


## 1. INTRODUCTION

There are many challenges in artefact detection in endoscopic images. Analysis of the dataset reveals two significant problems. Firstly, there is a class imbalance problem. While artefacts such as specularity account for nearly 34 percent of all detections, instrument class accounts for only 1.7 percent. Three classes (specularity, artifact, & bubbles), in total, account for 82 percent of all bounding boxes. Secondly, there is a scale imbalance problem. Various bounding boxes cover almost the entire frame & different bounding boxes only have very few pixels. Hence, the object detection algorithms' parameters should be chosen carefully in light of these observations to detect both small & large objects. We adopted an approach based on an ensemble of object detectors. Despite being slower, we mainly focused on two-stage networks due to their ability to detect small & very close objects & used Faster R-CNN & Cascade R-CNN. Also, we used a single-stage detector, RetinaNet, as a complementary model in the ensemble.

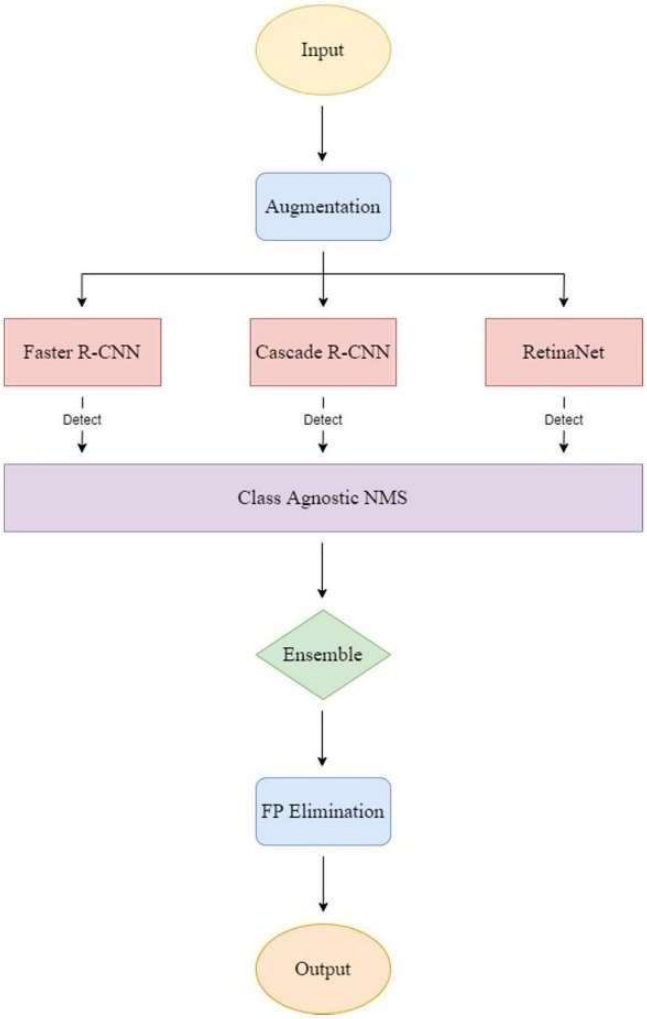


*Fig. 1: Flowchart of the Proposed Method*

## 2. PROPOSED METHOD

The flowchart of the proposed approach is given in Fig. 1. We use three base models. These base models' outputs are then fed into a class agnostic non-maximum suppression algorithm independently before combining the results through an ensemble model. Then a false-positive elimination is applied to the output of the ensemble model. In the remainder of this section, we describe these steps in more detail.

### 2.1 Base Models

We use two two-stage models: Faster R-CNN, Cascade R-CNN & one single-stage model, RetinaNet, as base models. Examination of the previous studies in this

domain reveals that feature pyramid network (FPN) & ResNet architectures achieve promising results. Therefore, these networks have been selected as the basis for our models.

The first model is based on Faster R-CNN & uses FPN as a backbone. Although FPNs are compute & memory intensive, they are good at extracting features at different scales. Since the dataset consists of objects in various sizes, FPNs are an essential element of the proposed network. We used a ResNet50 model with FPN as the backbone of this model. Standard convolutional & fully connected heads have been used for box predictions.

The second model is the Cascade R-CNN. While it is a similar model to Faster R-CNN, it is claimed to alleviate overfitting at training. Cascade R-CNN consists of consecutive detectors which are trained sequentially with increasing intersection-over-union (IoU) thresholds. This architecture is reported to be more selective against close false positives. Again, we used a ResNet50 model with FPN as a backbone.

In addition to these two-stage object detectors, we trained & used a RetinaNet as our third model. RetinaNet is a single-stage method &, as such, doesn't use a region proposal network. It has one backbone network that extracts features & two sub-networks for object classification & bounding box regression. An essential difference of this network from other single-stage networks (e.g., YOLO, SSD) is the use of focal loss. Focal loss is an extension to cross-entropy loss that puts a focus on sparse hard examples. It changes the weight of loss according to the performance of the model on different examples.

### 2.2 Class-Agnostic Non-Maximum Suppression (NMS)

In the original Faster R-CNN architecture, NMS operation is performed on each class independently. Yet, these architectures are generally designed considering non-medical datasets such as COCO or PASCAL VOC, which have high overlap ratios among the bounding boxes of different classes.

However, it isn't expected to have frequent overlaps between different objects in the endoscopic images. Therefore, we propose a class-agnostic procedure where the model predictions are passed through the NMS process together for all classes. As a consequence of this process, if multiple models with high IoU detect an artefact, the lower confidence scores are eliminated. A threshold of 0.4 IoU has been used to perform this class-agnostic NMS step.

### 2.3 Ensemble of Models

Two different ensemble methods, affirmative & consensus, have been used. In the affirmative method, the outputs of other models are merged, & NMS operation is applied to the result. It can be regarded as the union of all bounding boxes. In the consensus method, only the bounding boxes for which most of the models agree are kept. This method is similar to the ensemble of models in classification problems.

### 2.4 False-Positive (FP) Elimination

Although class-agnostic NMS discards the bounding boxes with high IoU with other bounding boxes in the detector network, the IoU threshold (0.4) might still be too high for the same class types. For example, if the intersection of two bubble bounding boxes has a very low probability, but the model predicts bounding boxes that have high IoU, it implies that there is redundancy & one of them should be removed. Therefore, we examined the IoU histogram of each class individually & determined a class-specific threshold. When there are bounding boxes with higher IoU values than the point, the ones having lower confidence scores are removed. Thresholds are determined according to the 1.5 interquartile range (IQR) above the 3rd quartile. Thresholds for elimination are given

| Class | Threshold | Class | Threshold |
|---|---|---|---|
| Specularity | 0.13 | Contrast | 0.19 |
| Saturation | 0.21 | Bubbles | 0.12 |
| Instrument | 0.24 | Blur | 0.4 |
| Artifact | 0.17 | Blood | 0.11 |

*Table 1: IoU thresholds for false-positive elimination*

in Table 1. This process is applied after the ensemble operation. Fig. 2 demonstrates the effect of this step.

## 3. RESULTS & DISCUSSIONS

According to the results in Table 2, while the individual networks have very similar mAP values, the Faster R-CNN model has a higher mIoU. The affirmative ensemble gives the highest mAP score as expected because some true positives missed by a model can be detected by the other models. On the other hand, a higher number of false positives are generated, which adversely affects its mIoU score. The consensus ensemble has the highest mIoU value among the methods, not utilizing FP elimination. Although class-agnostic NMS & FP reduction steps decrease the mAP values marginally, they eliminate many false-positives & give higher mIoU scores, resulting in a more balanced mAP & mIoU scores. Different score metrics are used for different object detection tasks. In this work, we have used post-processing techniques to have a balanced mAP & mIoU scores.

The highest scores are obtained using the consensus ensemble of the classifiers, which were passed through a class-agnostic NMS, & FP reduction as the final step.

Object detectors are generic & they aren't developed considering the domain-specific challenges. These networks also have many internal parameters & these parameters need to be tuned for the particular application. Hence, it isn't sufficient to use more advanced models & a comprehensive understanding of the characteristics of the data is of the essence.

To integrate the domain knowledge into detection architecture, we have qualitatively observed that some classes, such as specularity & saturation, have bounding boxes overlapping with each other. While removing the one that has less confidence seems to be a solution, this isn't ideal since, in several cases, the one with less confidence is the true class. Therefore, specific algorithms should be included in the detection framework to tackle this problem.

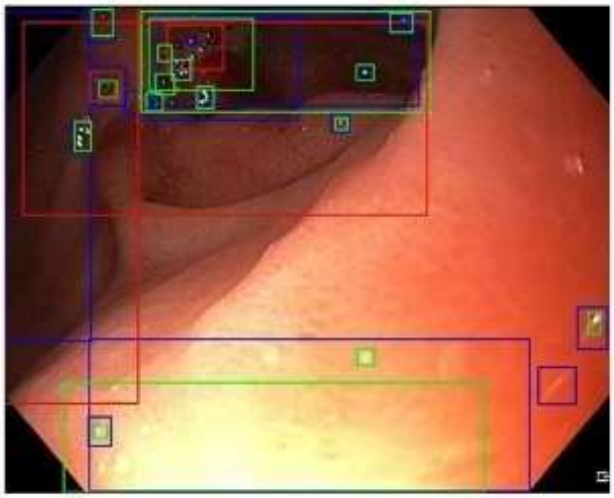

Fig. 2 Blue: Ground-truth bounding boxes.
Red: Bounding boxes eliminated after the FP reduction step.
Green: Remaining predicted boxes after elimination.

## 4. CONCLUSIONS

In this study, we have trained three different object detectors for endoscopic artefact detection. We have used ensemble

| Method | Without Class-Agnostic NMS | | With Class-Agnostic NMS | |
|---|---|---|---|---|
| | mAP | mIoU | mAP | mIoU |
| **Faster R-CNN with FPN** | 45.76 | 40.68 | 44.15 | 42.85 |
| **Cascade R-CNN with FPN** | 46.02 | 32.37 | 44.01 | 35.04 |
| **RetinaNet** | 45.10 | 36.45 | 43.92 | 41.22 |
| **Ensemble (A)** | **47.92** | 26.03 | 47.13 | 30.29 |
| **Ensemble (C)** | 47.30 | 42.90 | 45.97 | 45.20 |
| **Ensemble (A) with FP elimination** | 46.93 | 32.22 | 46.55 | 34.26 |
| **Ensemble (C) with FP elimination** | 46.87 | 44.66 | 45.72 | **45.92** |

***Table 2:*** *Experimental Results*

techniques to utilize all three individual networks. Applying a class-agnostic NMS to each of them independently resulted in a better trade-off between mAP & mIoU scores. As a final step, FP elimination is used, which resulted in more robust results.

In this work, we have focused on using lighter networks & taken the ensemble of weak classifiers approach. The use of lighter networks made the hyper-parameter tuning possible in feasible periods & allowed us to experiment with various network parameters. In the future, more sophisticated networks, such as ResNeXt or ResNet152, which require more time to train & tune parameters, could also be investigated.

## 5. INDUSTRY APPLICATIONS

The model can localize bounding boxes, predict class labels & pixel-wise segmentation of 8 different artefact classes for given frames & clinical endoscopy video clips. Existing endoscopy workflows detect only one artefact class, which is insufficient to obtain high-quality frame restoration. In general, the same video frame can be corrupted with multiple artefacts, e.g., motion blur, specular reflections, & low contrast can be present in the same frame. Further, not all artefact types contaminate the frame equally. So, unless multiple artefacts present in the frame are known with their precise spatial location, clinically relevant frame restoration quality can't be guaranteed.

Another advantage of such detection is that frame quality assessments can be guided to minimize the number of frames that get discarded during automated video analysis. It can be used for precise detection of endoscopy artefacts, which is essential for high-quality endoscopic frame restoration & crucial for realizing reliable computer-assisted endoscopy tools for improved patient care.

## 6. ACKNOWLEDGEMENTS

We want to thank Professor G N Pillai Sir and Electrical Engineering Department, Indian Institute of Technology, Roorkee, to provide the workstation and GPUs used in this work.

## 7. REFERENCES

1. **Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn:** Delving into high-quality object detection. In Proceedings of the IEEE conference on computer vision & pattern recognition, pages 6154–6162, 2018.

2. **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.

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

4. **Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, & Piotr Doll' ar.** Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision, pages 2980–2988, 2017.

5. **Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun.** Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision & pattern recognition, pages 770–778, 2016.

6. **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, & Jens Rittscher.** An objective comparison of detection & segmentation algorithms for artefacts in clinical endoscopy. Scientific Reports, 10, 2020.