International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

Detecting Multi-Class Artifacts in Endoscopic Images using YOLOv3

N.Kirthika, B.Sargunam

Abstract*: *Endoscopy is a regular, clinical non-surgical procedure to examine hollow organs like esophagus, stomach, intestine, etc., Nowadays, due to rapid technological development and miniaturization of hardware endoscopy can be performed in the respiratory tract, urinary tract, female reproductive tract and joints apart from gastrointestinal (GI) tract. The need for GI endoscopy is to examine, and diagnose ailments like cancers, polyps, and assist cauterizing bleeding vessels. The organ nature creates lots of artifacts in the imaged tissue such as saturation, specularity, blur, contrast, bubbles, and debris which causes significant challenges in quantitative analysis. Thus, combining state-of-the-art deep learning object detection algorithms like YOLOv3 with endoscopy leads to efficient and accurate localization and categorization of different imaging artifacts. This paper presents a detailed implementation of YOLOv3 in detecting endoscopic artifacts. Intensive training in a GPU enabled environment (Google COLAB) was carried out. The experimental results of the algorithm achieve mAP% of 55.18 for identifying the artifacts in endoscopic images with a prediction time of 0.0179 seconds on test images.

Keywords: Artifact detection, Deep learning, Endoscopy, YOLOv3.

I. INTRODUCTION

Object detection using machine learning and deep learning is a popular area of research. The applicability of computer vision algorithms has extended to all fields of manufacturing, autonomous driving, space applications, medicine, etc. The challenge lies behind the speed and accuracy in every domain. The human brain accurately performs object detection tasks at an incredibly fast rate. It can detect, localize, and classify objects in the scene. All object detection algorithm proposed so far aims at fulfilling the same task with high accuracy and speed but humans do it with ease. The algorithm faces lots of challenges like limited dataset, speed, multiple spatial scales, different aspect ratios, class imbalance, etc., Despite numerous challenges many authors have proposed different algorithms for object detection in various applications [1] [2] [3].

Manuscript received on May 18, 2020.
Revised Manuscript received on May 27, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
N.Kirthika*, Research scholar, Department of Electronics and Communication Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India. Email: prof.kirthika@gmail.com.
Dr.B.Sargunam, Associate Professor & Head, Department of Electronics and Communication Engineering, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India. Email: sargunam_ece@avinuty.ac.in.

The approach deeply rooted from sliding – window methods [4] till object detection using Deep Convolutional Neural Network in trend like Faster R-CNN (Faster Region based Convolutional Neural Network) [5], Mask R-CNN [6] etc., CNN (Convolutional Neural Network) provides solutions to many computer vision problems with a decreased error rate, but at the cost of computation power and speed. Multi-class endoscopic artifact has its importance where in the past lot of algorithms are proposed to detect only specific artifacts [7] [8] in endoscopic images, but in practical conditions, one or more of artifacts may appear in the same frame for instance bubbles, contrast and specularity may appear in the same frame thus making the detection quite challenging. Endoscopic images have sparse features and do not possess a geometric structure. Endoscopists may switch to different modalities like normal bright-field (NBF), Narrowband imaging (NBI) etc., for highlighting required features, and also the analysis may vary between patients on the same modality. Thus, algorithms developed must be robust in detecting multiple artifacts under different modalities across different patients.

II. RELATED WORK

In recent days, with the advent of deep learning techniques, object detection algorithms have taken a new big step. Object detecting algorithms proposed in the past include two-stage detectors like R-CNN (Region based CNN), Fast R-CNN [9], Faster R-CNN [5], Mask-R-CNN [6], and single-stage detectors like YOLO (You Only Look Once) [10] and SSD (Single Shot Detector) [11]. These algorithms used different backbones for feature extraction. Bouget D et al. [3] presented a paper reviewing algorithms for detection and tracking of surgical tools, the authors quote that reliable accuracy is not attained due to lack of established dataset and non-availability of standard performance assessment where the availability could bring in faster improvement in the technology. Sharib Ali et al. [12] proposed an end to end deep learning framework which can precisely detect artifacts and restore mildly corrupted endoscopic video frames. The authors have investigated Faster R-CNN, RetinaNet, and proposed YOLOv3-SPP (YOLOv3-Spatial Pyramid Pooling) for artifact detection in endoscopic video. Significant improvements in quality metrics were proved for artifacts like specularity, contrast, and others but efficient improvement in PSNR (peak Signal to Noise ratio) and structural similarity for saturation and blur.

Zhang and Xie [13] proposed cascaded R-CNN by modifying Faster R-CNN and combining it with the Feature Pyramid Network (FPN). The proposed work improved IoU (Intersection over Union) between candidate output and ground truth and the authors used a phased approach to gradually increase the threshold values of IoU during the training of the cascaded R-CNN framework. Xiaohong W Gao and Yu Qian [14] utilized the existing deep learning frameworks like Fast R-CNN-nas, YOLOv3 with different thresholds (0.1 and 0.25), Fast R-CNN with resNet-101 backbone for multi-class artifact detection; and derived results stating Fast R-CNN with resNet-101 backbone performs better. Seiryo Watanabe [15] proposed an algorithm combining YOLOv3 and Mask R-CNN utilizing the Mask R-CNN framework for detecting artifacts like instruments, specularity, saturation, bubbles, and general imaging artifacts because of its clear boundaries, but for artifacts with unclear boundaries like blur and contrast YOLOv3 was preferred. The authors [12] [13] [14] [15] concentrated on segmentation of artifacts too.

In this paper YOLOv3 using the Darknet framework is used for detecting all 7 artifacts in endoscopic images as it retains the spatial relation of objects concerning the background. This algorithm addresses the detection and localization of endoscopic imaging artifacts with the publicly available EAD2019 dataset [16]. The YOLOv3 outperforms in speed when compared to existing object detection algorithms on standard open resource datasets [17] thus making the algorithm reliable for our application. The structure of the remaining article presented is as follows. Section III introduces YOLOv3 with its feature extractor, Section IV discusses on experimental setup followed by section V focuses on results and analysis and section VI concludes the work.

III. NETWORK ARCHITECTURE OF YOLOV3

YOLOv3 is a state-of-the-art object detection algorithm [18]. YOLO divides the whole image into different regions called anchors and draws bounding boxes using K means clustering with each bounding box having a confidence score calculated using logistic regression. There are 9 anchor boxes as shown in Fig. 1 in total which are applied to predict output at three different scales. The distribution of anchors to different scales is the same as COCO dataset.

```
anchors =  13, 17,  30, 34,  52, 69, 117, 77,  78,156, 228,133, 149,235, 268,283
, 394,384
```

Fig. 1.YOLOv3 anchors

YOLOv3 assigns only one bounding box for each ground truth. The confidence score reflects the possibility of the presence of an object inside the bounding box which is calculated using the formula (1).

Class Confidence score = Box confidence score * conditional class probability (1)

The algorithm predicts class probabilities based on multi-label classification. The box with a high confidence score is considered to have an object in it. The loss function is the weighted sum of classification loss, confidence Loss, and localization loss as shown in (2). The localization attributes use mean square error while classification and confidence loss use binary cross-entropy.

Total loss = Localization loss + Classification loss + Confidence loss (2)

Darknet 53, a Convolutional Neural Network used as a feature extractor (also called as the backbone). It has 53 convolutional layers hence the name Darknet53(Refer Fig. 2). It is a pre-trained network trained to classify images of imageNet dataset of 1000 categories.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 2.Darknet 53 Image source[18]

Three different scales of YOLO layers (1/8, 1/16, and 1/32) are used to detect objects of different scales. The final feature map, i.e., the output is 1/32 times smaller than the input image in terms of spatial resolution. The overall architecture of YOLOv3 is shown in Fig. 3.

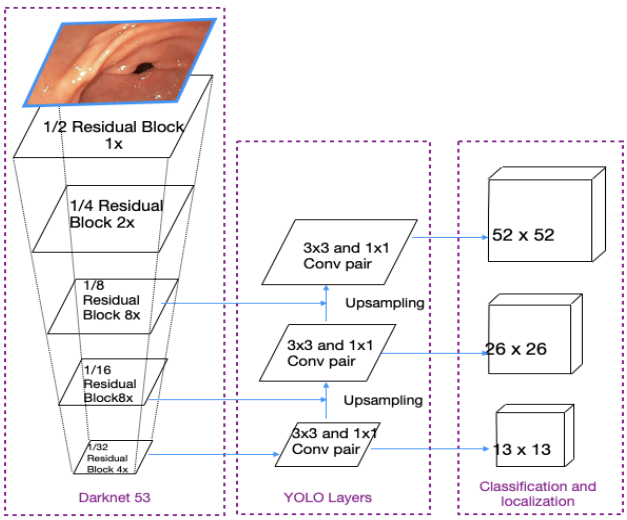

Fig. 3.YOLOv3 Architecture

This network has proven to be efficient in terms of speed than other feature extractors like ResNet-101 and ResNet-152 on image Net dataset and achieved the highest floating-point operations/second (BFLOPS/S) [18] because it looks at the whole image using a single neural network wherein other algorithms like R-CNN needs numerous single images to process.

In this paper pre-trained weights [19] which were originally trained to classify images from imageNet dataset is used as initial weights just to avoid training from scratch. This pre-trained feature extractor gives weights only for the convolutional layers and not to the fully connected layers. Since its inception YOLO has progressively developed from YOLOv1 to YOLOv3 and stepping into YOLOv4 which comes into the life of many applications soon.

IV. EXPERIMENTAL SETUP

A. *Data set*

The EAD2019 dataset [16] is a first-ever publicly available dataset for endoscopic artifact detection and segmentation. The images are compiled from 6 different centers worldwide. It is a multi-tissue, multi-modal, and also composed of people from different countries. The artifact detection dataset covers 7 different classes as shown in Table I. The implemented algorithm was trained and tested using the EAD2019 dataset.

Table- I: Classes in EAD2019 dataset

Class Number	Class Name	Source of Artifact
Class 0	Specularity	Mirror-like surface reflection
Class 1	Saturation	Overexposure to light
Class 2	Artifact	Includes general imaging artifacts, debris etc.,
Class 3	Blur	Due to hand movements and fast camera motion
Class 4	Contrast	Due to underexpose and occlusions
Class5	Bubbles	Water bubbles present in the track distorting the underlying tissue appearance
Class 6	Instrument	Instruments falling within imaging area

B. *Data Augmentation*

The data augmentation technique has become a staple technique for deep learning algorithms that are data thirsty. We use different augmentation techniques like varying hue, saturation, exposure, and flipping to increase the data samples, hence the algorithm can perform better and avoids overfitting.

C. *Training*

The datasets were divided into training, validation and testing sets. Training dataset contains 80% of the images and the rest with 10% of images for each validation and testing. Intensive training was carried on using high-performance GPU (Graphics Processing Unit) from Google Co-laboratory. The training parameters were set up along with data augmentation initializations. The total number of classes to be detected is 7 and thus classes are set to 7 in all 3 Yolo layers in the configuration file, correspondingly the number of filters must be changed with respect to the formula in (3).

$$\text{Filters} = (\text{Class} + 5) * 3 \quad (3)$$

Therefore, the filter parameter must be set to 36 in each convolutional layer that precedes YOLO layers [20]. Table II shows the input parameter set for training YOLOv3.

Table-II: YOLOv3 Input Parameters

Parameter Name	Parameter Values
Image Size(in pixels)	416 x 416 (Height x Width)
Batch Size	64
Subdivisions	16
Initial Learning rate	0.001
max batches = (2000 * class)[6]	14000

Darknet Neural Network is used for artifact detection. The Darknet Neural Network is an open-source framework written using languages like C and CUDA. It supports both GPU and CPU computations. The Darknet was cloned from their repository [20]. It is programmed to save weight for every 1,000 iterations. The training is continued until it reaches a minimum average loss. After around 50,000 iterations average loss is found to be flattened. The graph in Fig.4 below is plotted between the number of iterations and the average loss incurred. The iterations can be stopped once the average loss is no longer reducing or the loss has reached 0.05 for a small dataset and 3.0 for bigger datasets [20]. At the end of 50,000 iterations 32,00,000 images were trained.

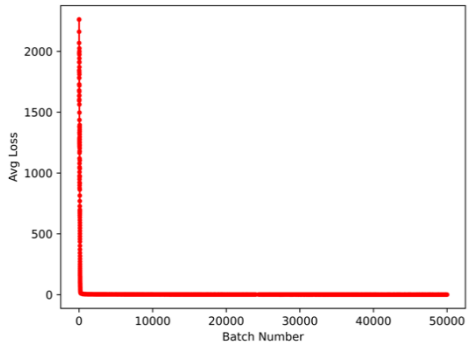

Fig. 4.Iterations(Batch number) vs Average loss plot

D. *Evaluation Indicators*

Deep learning models that are used to solve problems are generally evaluated using the validation/test set. Common performance metrics for an object detection algorithm include precision, recall, accuracy, etc., The choice of selecting the metrics depends upon the dataset and nature of the application. The most common metric used to evaluate the performance is mean average precision(mAP). This metric is used when prediction and localization of objects along with class is essential. Any performance metric used in an object detection algorithm is compared with the

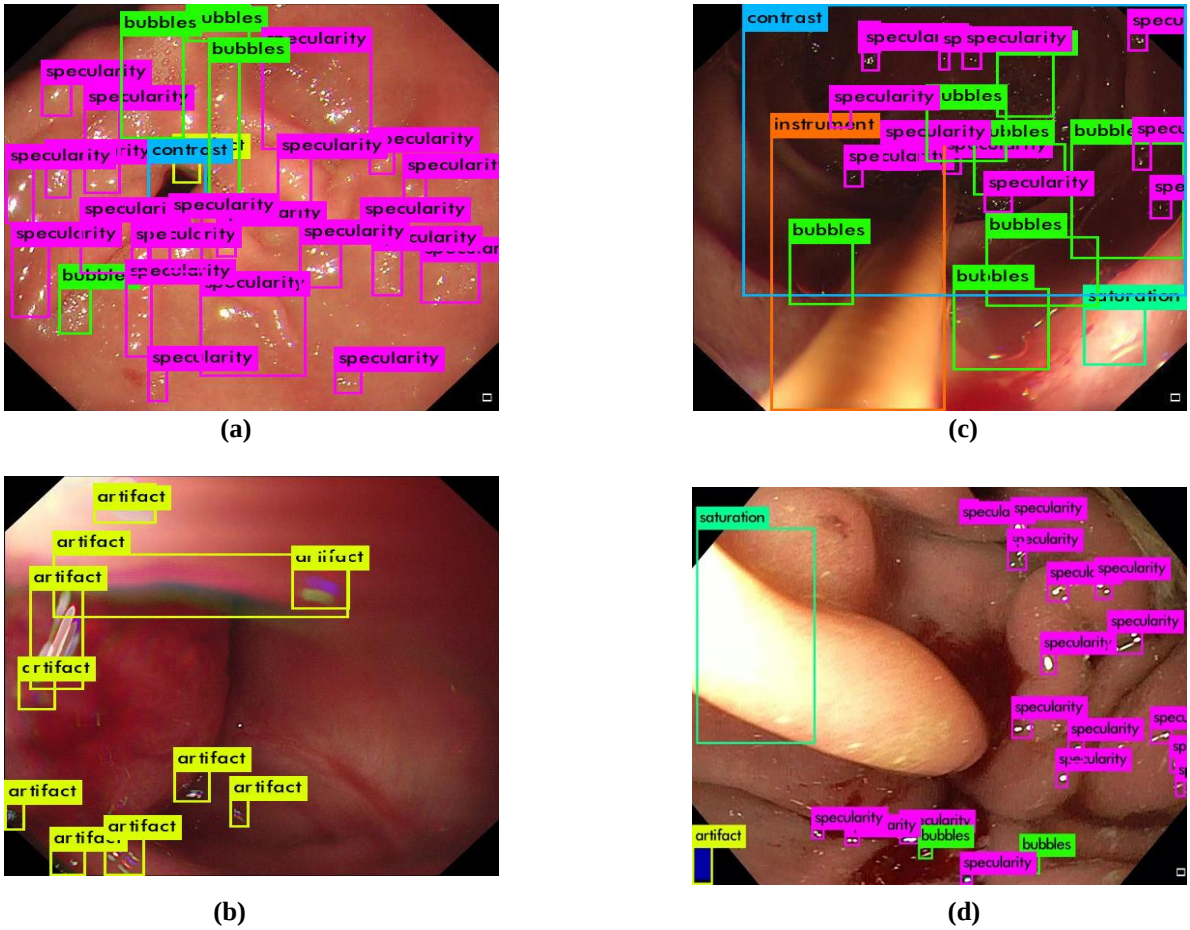

Fig. 5.Detection of multiple overlapping artifacts using YOLOv3

ground truth data. Where the ground truth data contain the image, bounding boxes for the object to be predicted by the algorithm and the class to which the object belongs to. The ground truth data must be made available for training, validation and test sets.

To calculate mAP first IoU must be calculated using (4). IoU is a ratio between the intersection of ground truth and predicted bounding boxes and union of ground truth and predicted bounding boxes.

$$IoU = \frac{A \cup B}{A \cap B} \tag{4}$$

IoU is used to get true and false positives. A threshold of 0.5 is set to classify the same. False negatives are a measurement calculated by detecting the number of objects that the algorithm missed out, which in turn used to calculate recall. Precision and recall can be calculated using the formula shown in (5) & (6).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

where TP represents True Positive (Positive samples predicted as positive samples), FP represents False Positive (Negative samples predicted as positive samples) and FN stands for false negative (positive samples predicted as negative samples). The next metric to be calculated is the Average Precision (AP) which is obtained by intersecting coordinate axis and the precision-recall curve at recall values (r_1,r_2,…). Equation (7) is used to calculate the Average precision (AP).

$$AP = \sum_i (r_{n+1} - r_n) p_{interp}(r_{n+1}) \tag{7}$$

where p_{interp}= max p(r) Finally, mAP is the mean of all average precisions calculated using the formula in (8).

$$mAP = \frac{1}{N} \sum_i AP_i \tag{8}$$

This metric was popularized in PASCALVOC challenge [17].

V. RESULTS AND ANALYSIS

The experimental results of YOLOv3 are shown in Fig. 5. It shows the bounding box obtained using YOLOv3 for multi-class artifact detection on the test set. Artifacts like instruments can be determined with ease, but artifacts like bubbles and saturation are densely distributed in most of the frames and also as stated earlier two or more artifacts are present in the same frame the algorithm is trained such that it can detect all the artifacts with better accuracy in spite of overlapping bounding boxes of different artifacts which is evident in Fig. 5(a) to 5(d).

Fig. 5(b) Shows general imaging artifact being detected with class name artifact. The algorithm took 17.939 milliseconds to detect artifacts in test images in GPU enabled environment. mAP (Mean Average Precision) score for every 1,000 iterations is plotted. Higher the mAP is better [20]. It is found from the graph that mAP% is higher at 49,000 iterations. The mAP calculations are performed over the validation dataset. Thus, a weight file of 49,000th iteration is considered for detections and performance analysis. The detections are done with the test dataset. The graph plotted between the iterations and mAP% scores obtained on the validation dataset is shown below in Fig. 6.

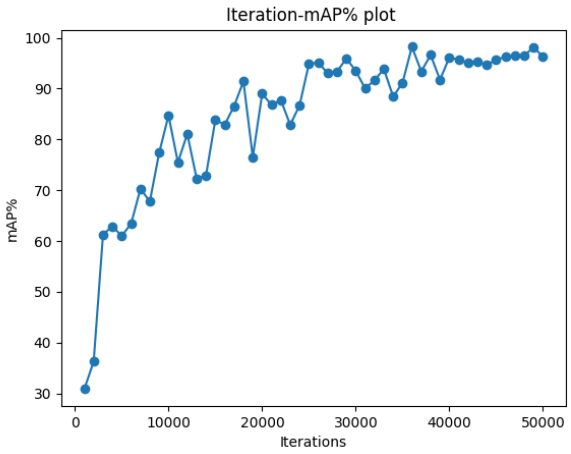

Fig. 6.Iteration Vs mAP% plot

The performance of YOLOv3 is tabulated in Table III. The performance greatly relies on the threshold of the bounding boxes used for object detection. An increase or decrease in the threshold will directly lead to changes in the number of bounding box proposals. Thus, mAP is calculated with different thresholds and the results are tabulated. It is evident from the results that 50% of threshold holds good for the application artifact detection with precision and recall values to be 0.62 & 0.50.

Table- III: mAP Scores of YOLOv3

Method	mAP%@ IoU 50	mAP% @ IoU 75
YOLOv3	55.18	30.83

Precision detection of each class is shown in Table IV

Table- IV: AP scores for each class

Class	Name	AP Scores
Class 0	specularity	44.43
Class1	saturation	46.41
Class 2	artifact	40.91
Class 3	blur	88.89
Class 4	contrast	51.96
Class 5	bubbles	13.63
Class 6	instrument	100.00

The performance of the algorithm in detecting various artifacts are proved with AP scores. The difference in percentage between classes is because of the imbalance identified in the class distribution of the training dataset. The overall detection has achieved good results when compared to mAP scores and prediction time of two-stage detectors [21].

VI. CONCLUSION

We have demonstrated the usage of YOLOv3 for localization and detection of 7 commonly occurring artifacts in endoscopic images. The images from the EAD2019 dataset are used for training, validation, and testing. The experimental setup with aforesaid input parameters results in mAP scores of 55.18% which shows the efficiency of the algorithm in predicting artifacts amidst of irregular object structure and very small dimensions. The prediction time of 17.939 milliseconds is obtained which marks YOLOv3 to be the fastest algorithm again. Thus computer-assisted tools aided with such an object detection algorithm can help doctors in input processing and review/report preparation steps and as well assist fresh endoscopists.

VII. FUTURE WORKS

To improve IoU, anchor box dimensions can be modified so that tiny objects in the image such as specularity and bubbles can be more effectively detected with increased confidence.

ACKNOWLEDGMENT

The authors thank the Management, the Director and the Dean, School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women for providing computing facilities and necessary support for the conduct of experiments and investigations.

REFERENCES

1. H. Schneiderman and T. Kanade “A statistical method for 3D object detection applied to faces and cars,” in the proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol.1, pp. 746–751, 2000.
2. M. H. Yang, David Kriegman and Narendra Ahuja, “Detecting faces in images: a survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 34-58, 2002, DOI: 10.1109/34.982883.
3. Bouget D., Allan M., Stoyanov D., and Jannin P., “Vision-based and marker-less surgical tool detection and tracking: a review of the literature”, Medical Image Analysis, Vol. 35, pp-633-654, 2017. DOI: 10.1016/j.media.2016.09.003.
4. P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in the proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 511-518,2001, DOI: 10.1109/CVPR.2001.990517.
5. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,”IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149,2016. DOI: 10.1109/TPAMI.2016.2577031
K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” IEEE Transactions on Pattern Analysis and Machine intelligence, vol. 42, no. 2, pp. 386–397, 2020, DOI: 10.1109/TPAMI.2018.2844175.
6. Thomas Stehle, “Removal of specular reflections in endoscopic images” Acta Polytechnica: Journal of Advanced Engineering, Vol. 46, no.4, pp-32-36, 2006.

7. H. Liu, W.-S. Lu and Max Q.-H. Meng, "De-blurring wireless capsule endoscopy images by total variation minimization," in proceedings of 2011 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, Victoria, BC, 2011, pp. 102-106, DOI: 10.1109/PACRIM.2011.6032875.
8. R. Girshick, "Fast R-CNN," in Proceedings of IEEE International Conference on Computer Vision, pp. 1440–1448, 2015, DOI: 10.1109/ICCV.2015.169.
9. J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: unified, real-time object detection," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, DOI: 10.1109/CVPR.2016.91.
10. Wei Liu et al., "SSD: Single shot multibox detector," European Conference on computer vision (Lecture Notes Computer Science book series), ECCV 2016, vol. 9905 LNCS, pp. 21–37, 2016.
11. S. Ali et al., "A deep learning framework for quality assessment and restoration in video endoscopy," CEUR Workshop Proceedings vol. 2366, 2019, Available: http://arxiv.org/abs/1904.07073.
12. Yan-yi Zhang and Di Xie, "Detection and segmentation of multi-class artifacts in endoscopy," Journal of Zhejiang University Science B, vol. 20, no. 12, pp. 1014–1020, 2019.
13. Xiaohong W. Gao and Yu Qian, "Patch based deep learning approaches for artefact detection of endoscopic images," CEUR Workshop Proceedings, vol. 2366, paper-10, 2019.
14. Seiryo Watanabe, Shigeto Seno, and Hideo Matsuda, "DNN models and postprocessing thresholds for endoscopy artifact detection in practice," CEUR Workshop Proceedings, vol. 2366, paper-7, 2019.
15. S. Ali et al., Endoscopy Artifact Detection (EAD2019) challenge dataset, DOI:10.17632/C7FJBXCGJ9.1.
16. Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn and Andrew Zisserman, "The Pascal Visual Object Classes (VOC) challenge," in International Journal of Computer Vision, Vol. 88, pp. 303-338, 2010. DOI: https://doi.org/10.1007/s11263-009-0275-4.
17. J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement" pp. 1-6, 2018. arXiv:1804.02767v1.
18. J. Redmon, Darknet: Open source neural networks in c. Available: http://pjreddie.com/darknet/.
19. AlexyAB github repository, Available: https://github.com/AlexeyAB.
20. Qingtian Ning, Xu Zhao and Jingyi Wang, "Deep layer aggregation approaches for region segmentation of endoscopic images", CEUR Workshop Proceedings, vol. 2366, paper-8, 2019.

AUTHORS PROFILE

N.Kirthika received her Bachelor of Engineering degree in Electronics and Communication Engineering from Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore in 2009, and received a Master of Engineering degree from Anna University of Technology, Coimbatore in the year 2011. She is currently pursuing her full-time research at School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. Her area of interest includes High Speed VLSI Architectures, Image processing and Deep Learning algorithms.

Dr.B.Sargunam received her Bachelor of Engineering degree in Electronics and Communication Engineering from Karunya Institute of Technology, Coimbatore in the year 1994, Master's degree in Applied Electronics from PSG College of Technology, Coimbatore in the year 2008 and Doctor of Philosophy from Anna University, Chennai in the year 2015. She is currently working as Associate Professor and Head, Department of Electronics and Communication Engineering at School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore. Her research interest includes high-performance VLSI architectures and cryptography.