Multi-Object Detection and Localization of Artefacts in Endoscopy Images

Madhura Prakash M.
Research Scholar, Dept. of CSE,
BNM Institute of Technology, Bangalore

Krishnamurthy G.N., PhD
Principal
BNM Institute of Technology, Bangalore

ABSTRACT
In this era the Artificial Intelligence (AI) combined with computer vision techniques are seamlessly applied across various domains. The medical image analysis domain is also gaining the advantage of the AI solutions. The medical domain requires real-time analysis of the images and videos being generated for providing automatic assistance to the experts. Artefacts Artefact are image features that do not represent any original scene but occur due to a quirk of the modality itself. The presence of artefacts in Images or video frames poses a challenge for efficient analysis to extract the relevant information. The actual information in the image for understanding the given scene usually lies behind the presence of these artefacts. Examples of artefacts include distortion, blurring, occlusion by other objects and so on. The presence of these in an image must be identified and in the case of video analysis, the frames without the presence of with minimal presence of artefacts must be considered for analysis. Endoscopy is a procedure that involves both diagnosis and therapeutic solutions in various inner regions of human body. Analyzing the image data generated by this procedure using AI based solution can provide an assistance to the medical experts. The work focuses on providing deep-learning based result based on the standard YOLO V3 model for artefact detection and localization on the endoscopy frames has been proposed. The proposed model has achieved a mean average precision (mAP) of 0.76 and an Intersection of Union (IoU) of 0.63 by training the model on the images from the widely available Endoscopy Artefact Detection (EAD) dataset.

General Terms
Computer vision, Artificial intelligence, Endoscopy artefacts, Object Detection and Localization

Keywords
Endoscopy, Deep Learning, YOLO V3

1. INTRODUCTION
Endoscopy is a medical procedure that is used to look inside the hollow organs of human beings. This procedure aids both during the diagnosis events where the presence of any abnormalities can be assessed and also during the therapy where any surgery is performed. This process produces video data and the frames thus generated can either be assessed during real-time of the procedure being performed or can be assessed at a later stage. Careful manual assessment of these frames requires a lot of focus time by the experts. Computer vision based artificial intelligence solutions aid the experts by extracting the relevant portions of the video or relevant portions of an individual frame that require expert's attention and focus to make critical decision. One of the major challenges in any image analysis is the presence of unavoidable artefacts. These artefacts range from blurred image, presence of distortion, reflections due to light variation and different modalities to presence of occluded objects hiding the relevant portion of the frame. These artefacts make the image analysis a difficult task. The first step to overcome this difficulty is to detect the exact presence of these artefacts.

The number of artefacts that can be present in the Endoscopy frames are varied and numerous as compared to the number of artefacts that can be present in images from other modalities. This is mainly due to suboptimal light reflection, tissue movements that cannot be controlled, organ shape that varies and texture of the pathological structure, occlusions caused from body fluids and body waste. The imaging phenomena commonly causes overexposure and underexposure of image areas due to changes in illumination and organ topology. The blurriness, variations in saturation and contrast also occur due to the unsteady hand movements by endoscopist and irrepressible local organ movements. The light reflection from smooth surfaces of organs also makes it challenging.

It is necessary and important to have a solution that can detect the presence of all categories of artefacts to deal with them in an appropriate way during the analysis procedure. Detection of the artefacts in endoscopy frames are challenging because of reasons including: difference in tissue appearance under different modalities of endoscopy, diverse range of artefacts that have very less inter and intra class variations, overlap, varied and multiple occurrences of artefacts in a single frame. Ignoring the presence of these artefacts can cause a huge setback in the quality of the image analysis thus follows. An artefact detection procedure can also contribute in assessing the quality of the endoscopy procedure being performed and also can aid the physician by providing feedback for improvement. In this work the images from the Endoscopy Artefact Detection (EAD) dataset have been used to train one of the best performing neural networks in object detection and localization namely, YOLO V3.

2. EXISTING WORKS
The researchers in [2] have proposed an end-to-end solution of detection of varied classes of artefact presence in the endoscopy video frames and also to restore the original frames by the removal of these artefacts. The frame restoration techniques were based on the blind deburring and saturation removal which used the cyclic generative adversarial network (CGAN). Specular reflection and other miscellaneous artefacts detected by the YOLO V3 model were removed by a combination of image inpainting and CGAN based approach. 25% of the frames were restored and a mAP of 68.7% has been achieved.

An integrated approach was followed by the authors in [3] for detection and localization of artefacts. The proposed architecture was based on cascaded Region based Convolution Neural Network (RCNN) that received input from ResNet-101 as backbone and Feature Pyramid Network

(FPN) as neck. The method has achieved an F2 score of 0.6225 on the segmentation set.

A comparative analysis of all the solutions submitted to the initial challenge is presented in detail by the authors in [4]. This work helps in reviewing the state-of-the-art designs that have been used by the researches in proposing solutions to all the folds of the challenge task like detection and localization of artefacts, segmentation of the artefacts and also to assess the generalizability of the proposed model's performance on unobserved similar frames.

An analysis of all the solutions provided to EndoCV2020 challenge [5] has been given. The tasks both artefact and disease detection and segmentation have been considered and architecture solutions for these tasks have been analyzed in detail. Faster R-CNN is the commonly used model for detection purposes and U-Net based solution are common for segmentation. The work in [6] follows a combined approach based on R-CNN and FPN module. The images used are based on the original training data and translated images using cyclic GAN. The authors in [7] have used data augmentation techniques to improve model's generalizability. "EfficientDet-D2" has been used with transfer learning the model's weights and th parameters have been fine-tuned during training to detect the artefacts in endoscopy. The authors in [8] have used RetinaNet and Faster-RCNN models to locate endoscopic artefacts, and Deeplab v3 and U-Net models for the segmenting endoscopic artefacts. The authors in [8] have proposed an ensemble-based technique for dealing with high variability in the data for artefact detection. Here the models have been trained on different optimization plateaus.

3. METHODOLOGY

3.1 Material

The proposed method utilizes the data from the widely available EAD 2019 dataset [1]. The dataset has images collected from 6 different hospitals worldwide. The endoscopic video dataset is a combination of multi-tissue data including images from gastroscopy, cystoscopy, gastroesophageal and colonoscopy. These images also represent multimodality by consisting images from white light, fluorescence, and narrow band imaging. The dataset has been publicly released to assess the solutions on 3 challenges namely, artefact detection and localization, segmentation of artefacts and evaluating the model generalizability. This work focuses on the first of the tasks namely to detect and localize the artefacts.

A total of 2147 annotated frames are included in the training dataset. This constitutes images from 7 artifact classes namely Specularity, Saturation, artifact, contrast, blur, bubbles and instrument. The annotations provided in the training data are to be understood as: (i) blur - alternating from fast camera motion (ii) bubbles - water bubbles that alters appearance of the underlying tissue (iii) specularity - mirror-like surface reflection from smooth organs (iv) saturation – which represents overexposed bright pixel areas (v) contrast – that included low contrast areas from underexposure or occlusion (vi) artifact - assorted artifacts; e.g., chromatic aberration, debris, imaging artifacts (vii) instruments -presence of surgical instruments in case of therapeutic endoscopy procedure.

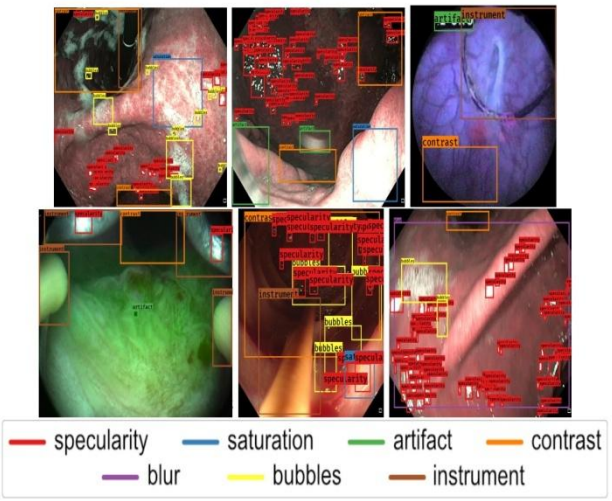

Fig 1: Example annotated images from the dataset [1]

3.2 Network Architecture

The proposed solution for multi-artefact detection and localization is based on the YOLO v3 architecture. YOLO (You Only Look Once) is a cohort of deep-learning based solutions designed for object detection. It is based on the idea that "A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance" [9]. YOLO v3 receives input as colored images of dimension 416*416. The network permits an input image into a convolution neural network (CNN) architecture. The last two layers of output are compressed resulting in an output volume from the architecture. Filters of a 19 x 19 grid returns 5 anchor boxes per grid. Each grid in the anchor box represents the number of classes that has to be detected. The bounding box will have six coordinates depicting the points (pc, bx, by, bh, bw,c). The output is a list of bounding boxes along with the recognized classes. An expansion of c into a 7-dimensional vector results in each bounding box. There an be an overlapping in the bounding boxes selected per grid. This can represent variance in class detection. Therefore, the IoU and Non-Max Suppression is followed to avoid picking overlapping boxes. The non-max suppression threshold is the technique that is incorporated to overcome the problem of detecting an artefact multiple times in a single frame. This ia achieved by considering the probabilities of each box and taking the anchor boxes with maximum probability. The boxes with close proximity are suppressed by taking into account boxes that do not have maximum occurrence of probabilities for a particular class of artefact. The representative diagram of the YOLO V3 architecture is given in figure 2.

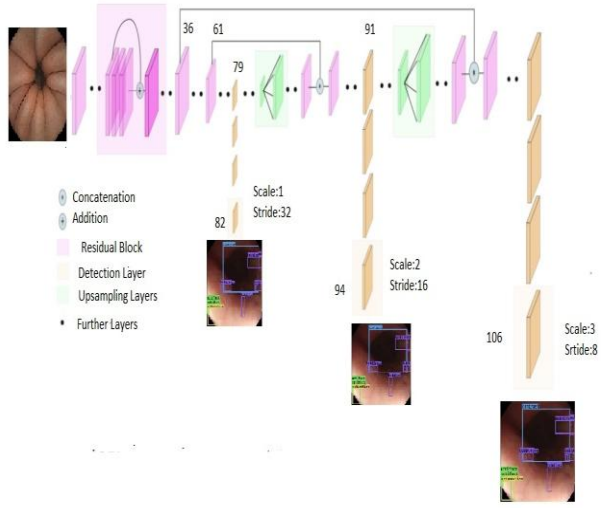

Fig 2: A representative diagram of YOLO V3 architecture

The DarkNet architecture originally consists of a 53 layered neural network which are pre-trained on the Imagenet data [10]. The YOLO v3 architecture uses a version of Darknet. In this case 53 more layers are stacked for the purpose of detecting artefacts in a considered frame. This has resulted in a 106 layer fully convolutional neural network architecture. The task of artefact detection a frame is accomplished by convolving kernels of shape 1 x 1 on the feature maps of at three different places in the network. The shape of the detection achieved by a convolving kernel is 1 x 1 x (B x (5 + C)). B represents the number of bounding boxes a grid on the feature map can predict. 5 accounts for the four bounding box attributes and the reaming one represents the confidence acore for a particular task. C indicates the number of classes of artefacts which is 7 in this scenario. A binary cross-entropy loss function is computed in each epoch for assessing the classification loss for each label. Logistic regression is used for object confidence and class predictions. The complete layer details of the network are a shown in the figure 3.

Layer	Filters size	Repeat	Output size
Image			416 × 416
Conv	32 3 × 3/1	1	416 × 416
Conv	64 3 × 3/2	1	208 × 208
Conv Conv Residual	32 1 × 1/1 64 3 × 3/1	Conv Conv Residual × 1	208 × 208 208 × 208 208 × 208
Conv	128 3 × 3/2	1	104 × 104
Conv Conv Residual	64 1 × 1/1 128 3 × 3/1	Conv Conv Residual × 2	104 × 104 104 × 104 104 × 104
Conv	256 3 × 3/2	1	52 × 52
Conv Conv Residual	128 1 × 1/1 256 3 × 3/1	Conv Conv Residual × 8	52 × 52 52 × 52 52 × 52
Conv	512 3 × 3/2	1	26 × 26
Conv Conv Residual	256 1 × 1/1 512 3 × 3/1	Conv Conv Residual × 8	26 × 26 26 × 26 26 × 26
Conv	1024 3 × 3/2	1	13 × 13
Conv Conv Residual	512 1 × 1/1 1024 3 × 3/1	Conv Conv Residual × 4	13 × 13 13 × 13 13 × 13

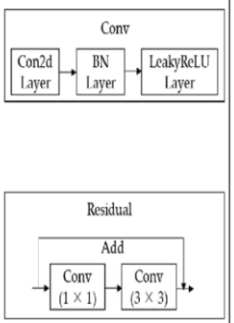

Fig 3: Specifics of YOLO V3 layer details in an illustrative diagram

The architecture of the YOLO model is comprised of only convolutional layers. This results in a fully convolutional network (FCN) structure that consists of 53 convolutional layers. The convolution layers are preceded by the batch normalization layer (BN). The activation function incorporated is the Leaky ReLU activation. This activation function being a better variant of the Rectified Linear Unit activation (ReLU) activation function overcomes the problem of dying gradient. [11] This function has a mean activation close to zero ensuring faster training speeds. This is achieved by having tiny slopes for negative values instead of having zero slope for all negative values. In the convolution layers of the architecture multiple filters are convolved on the frames. These filters produce a number of feature maps The network does not have a pooling layer. A convolutional layer with a stride of 2 is expended to down sample the feature maps. Pooling may cause the loss of low-level features and this down sampling helps to overcome the issue of low-level feature loss.

3.3 Proposed Solution

Image pre-processing is a general first step before consuming these images for training the model. The proposed solution uses a single step pre-processing of the images by down-sampling all the images to a dimension of 416 x 416 before feeding it into the model. The input to the model must be an RBG image with a resolution of 416 x 416 pixels. This pre-processing step which involves image resizing is done via aspect ratio aware function and by using the interpolation methods to keep the distortions minimal. The output produced by the model on a test frame will be a list of bounding boxes along denoting the recognized classes of artefacts present in the test frame. Each bounding box is represented by 6 numbers (pc, bx, by, bh, bw, c). The c value represents an 80-dimensional vector corresponding to the number of classes of artefacts to be detected in the frame, which is 7 in this case. Each of the bounding box is represented by a feature vector.

A set of of NumPy arrays will be returned as output by the application initially. The output display will be the shape of these NumPy arrays. The bounding boxes and class labels of each artefact present in the frame are predicted and depicted by these arrays. The values are present in an encoded format which has to be further processed. The NumPy arrays are then considered to decode the intermediary output namely the candidate bounding boxes and class predictions. The bounding boxes with class probabilities less than the threshold which is 0.3 are considered not to be confident in identifying the artefact and thus they are ignored.

A threshold of the number of bounding boxes in a test frame can be set during the application execution. A trial with a maximum of 200 bounding boxes is executed for the test frames. A function has to aid in performing translation of bounding box coordinates has been implemented. This function helps in plotting the original image and draw the bounding boxes on the test frame to depict the present of artefact in the frame. The redundant bounding boxes representing same artefact in a frame is removed by setting a threshold of 0.45 for non-max suppression [12]. Trials of the application has been conducted for both the presence of redundant bounding box and absence to consider the multiple occurrences of same artefact for example bubbles in a certain test frame.

The post-processing steps include the upscaling the bounding boxes to the original image. The coordinates of the bounding boxes are taken into account for rescaling the bounding box and to display the label of each artefact present in the frame. The labels are depicted on top of each box. This step is performed to display the bounding boxes along with the

recognized classes in the resulting output image. The model is able process approximately 23.6 images per second on a GPU based system

Since optimization is one of the major areas of concern in the deep learning model generation. An efficient optimization technique namely model weight pruning is also explored in this work. [13] In weight pruning the instead of considering full-precision for the floating-point values of the weights a half precision can be considered. The model can achieve a processing speed of 60.8 image/sec in a specified resolution of 416 x 416 pixel by pruning the weights of the model at half-precision (FP16) on a graphics processing unit (GPU). This helps in achieving a significant improvement on performance thereby helping in optimization. The initial solution of 19.6 image/sec using a 416 x 416 image can be achieved by using the model weights at full floating-point precision (FP32) on a GPU. [14] This is a significant trial of the proposed method to achieve better optimization in model training. This solution can also aid in generating models that are significantly lighter and thereby helping in edge deployment. The CUDA library solutions and optimized training solution contributes to the significant improvement in both mean average precision (mAP) and inference performance on GPUs.

4. RESULTS AND DISCUSSION

4.1 Outcome Metrics

The filters that convolve in the output layers of the proposed model predicts the bounding boxes depicting the occurrence of artefacts in the frame under consideration. The anchors organize all the boxes such that they are centred in the particular cell. After each epoch the predicted bounding boxes are compared against the ground truth boxes as given in the training set to compute the model loss at that point. [15] The loss computation in the architecture is based on (i) Coordinate loss — this occurs if the artefact is not exactly covered by the bounding box (ii) Objectness loss — this occurs if the intersection over union indicates that a wrong artefact has been identified by the bounding box as compared to its ground truth (iii) Classification loss — this is computed for each class of artefact and if the class prediction is reversed as compared to ground truth values.

The IoU (Intersection over Union) metric helps in assessing the quality of the bounding box generated by comparing the artefact prediction, with its ground truth box specified in the training set of the dataset. IoU values range from 0 (the bounding box totally misses the artefact detection) to 1.0 (a bounding box perfectly fits the artefact detection). For every spatial cell, for each artefact prediction in focus in that cell, the loss function aims at finding the bounding box with the best IoU with the artefact focused in that cell. This is a mechanism that help in distinguishing between the best boxes and all the predicted other boxes representing the artefacts in each spatial cell. The coordinate, objectness and classification loss computed for each batch, each epoch is used as a refence to fine tune the model's performance and gain better precision after multiple batches

The intersection over union (IoU) is a metric that supports to know if a particular bounding box depicts an artefact or not. The area of intersection and are of union of 2 bounding boxes are computer and the IoU is calculated according to the equation 1 by dividing the area of intersection between two bounding boxes by the area of union [16] For better prediction of the classes of artefacts a higher IoU is expected.

IoU= (Intersection Area) / (Union Area) ------- Eq 1

The mean average precision (mAP) is computed by starting with calculation the average precision (AP) for each class of the artefact presents in a test frame. The mean of the average precision all classes of artefacts in a test frame is the mAP. The mAP is computed as given in the equation 2.

mAP = 1/n ∑ K=1 to n APK --------- Eq 2

Where APK denotes the average precision of class of artefact and K and n is the number of classes which is 7 in this work. The model's iteration versus loss curve under various threshold of training is as represented in the figure 4.

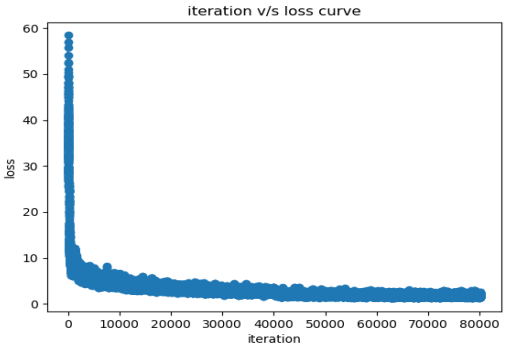

Fig 4: Iteration versus loss curve of model training

The Intersection over union (IoU) and the mean average precision (mAP) percentage of the model during the iterations of the several epochs are tabulated and the values are given in Table 1.

Table 1. IoU and mAP of the model during training iterations

Sl.No.	Iteration	IoU%	mAP%
1	1000	11.66	1.17
2	18000	20.56	10.73
3	36000	27.72	17.86
4	66000	30.89	14.92
5	80000	24.86	12.26

4.2 Outcomes

The 20 percent of validation images and the test images have been resized to 416*416 as per the requirement of the model. The output of the model is obtained as encoded candidate bounding boxes for the possible artefacts that can be present in an endoscopy frame and these bounding boxes are defined the framework of the anchor boxes represented in a frame.

Each one of the NumPy arrays representing the encoded bounding boxes is considered, one at a time by the implemented function these candidate bounding boxes and class predictions are decoded to denote the predict the artefact presence. The bounding boxes that do not fall within the threshold of the confident score to describe an artefact are ignored. A set of bounding boxes with the definition of its coordinates are returned by the application. The sample output image is given in figure 5 and figure 6.

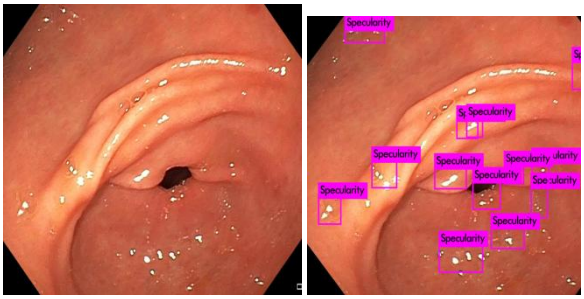

Fig 5: Few of the artefact detected from YOLO V3 architecture

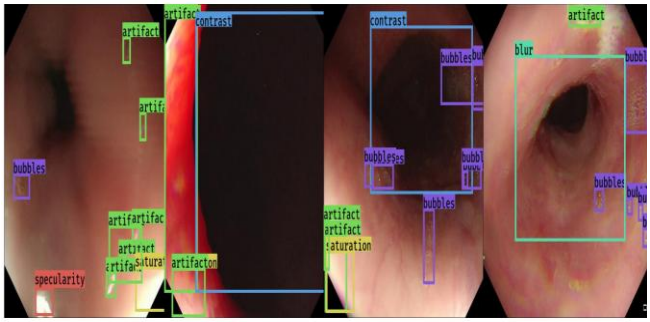

Fig 6: Another test image and the corresponding detection and annotations

5. CONCLUDING NOTES

In this work an artefact detection and localization solution has been provided based on the YOLO V3 architecture. YOLO V3 has proven to be one the state-of-the-art architectures in multi object detection and localization. The model has been trained and tested on the widely available EAD2019 dataset. The model has achieved a mean average precision (mAP) of 0.76 and an Intersection of Union (IoU) of 0.63.

Also, to bring in better optimization in the model training performance, the concept of pruning has been explored where the weights of the network has been pruned to half precision floating point values and thereby increasing the model training performance on the GPU and this has also helped in achieving better mean average precision.

In this work, only one of the tasks of the EAD2019 challenge has been addressed. Further 2 more tasks namely semantic segmentation of the artefacts in the frames and exploring and analyzing the model's generalizability by performing an ablation study is yet to be explored. A customized ensemble of state-of-the art models can be a best approach to achieve better precision and accuracy and also in developing an optimized solution.

6. REFERENCES

[1] Ali, Sharib; Zhou, Felix; Daul, Christian; Braden, Barbara; Bailey, Adam; East, James; Realdon, Stefano; Georges, Wagnieres; Loshchenov, Maxim; Blondel, Walter; Grisan, Enrico; Rittscher, Jens (2019), "Endoscopy Artefact Detection (EAD) Dataset", Mendeley Data, V1, doi: 10.17632/c7fjbxcgj9.1

[2] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James E. East, Xin Lu, Jens Rittscher, A deep learning framework for quality assessment and restoration in video endoscopy, Medical Image Analysis, Volume 68, 2021,101900, ISSN 1361-8415, https://doi.org/10.1016/j.media.2020.101900.

[3] Zhang YY, Xie D. Detection and segmentation of multi-class artifacts in endoscopy. J Zhejiang Univ Sci B. 2019 Dec.;20(12):1014-1020. doi: 10.1631/jzus.B1900340. PMID: 31749348; PMCID: PMC6885408.

[4] Ali S, Zhou F, Braden B, Bailey A, Yang S, Cheng G, Zhang P, Li X, Kayser M, Soberanis-Mukul RD, Albarqouni S, Wang X, Wang C, Watanabe S, Oksuz I, Ning Q, Yang S, Khan MA, Gao XW, Realdon S, Loshchenov M, Schnabel JA, East JE, Wagnieres G, Loschenov VB, Grisan E, Daul C, Blondel W, Rittscher J. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. Sci Rep. 2020 Feb 17;10(1):2748. doi: 10.1038/s41598-020-59413-5. PMID: 32066744; PMCID: PMC7026422.

[5] Sharib Ali, et al, Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis, Volume 70, 2021, 102002, ISSN 1361 8415,https://doi.org/10.1016/j.media.2021.102002.

[6] Yang, Suhui and Guanju Cheng. "ENDOSCOPIC ARTEFACT DETECTION AND SEGMENTATION WITH DEEP CONVOLUTIONAL NEURAL NETWORK." (2019).

[7] Yin, TK., Huang, KL., Chiu, SR. et al. Endoscopy Artefact Detection by Deep Transfer Learning of Baseline Models. J Digit Imaging (2022). https://doi.org/10.1007/s10278-022-00627-6

[8] Jadhav, Suyog & Bamba, Udbhav & Chavan, Arnav & Tiwari, Rishabh & Raj, Aryan. (2020). Multi-Plateau Ensemble for Endoscopic Artefact Segmentation and Detection.

[9] Redmon, Joseph & Farhadi, Ali. (2018). YOLOv3: An Incremental Improvement.

[10] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[11] Subramanian, Anand and Koushik Srivatsan. "Exploring Deep Learning Based Approaches for Endoscopic Artefact Detection and Segmentation." EndoCV@ISBI (2020).

[12] Oksuz, I., Clough, J. R., Clough, J. R., & Schnabel, J. A. (2019). Artefact detection in video endoscopy using retinanet and focal loss function. CEUR Workshop Proceedings, 2366. https://doi.org/http://ceur-ws.org/Vol-2366/

[13] Khan, M. A., & Choo, J. (2019). Multi-class artefact detection in video endoscopy via convolution neural networks. CEUR Workshop Proceedings, 2366.

[14] Sulik, L., Krejcar, O., Selamat, A., Mashinchi, R., Kuca, K. (2015). Determining of Blood Artefacts in Endoscopic Images Using a Software Analysis. In: Núñez, M., Nguyen, N., Camacho, D., Trawiński, B. (eds) Computational Collective Intelligence. Lecture Notes in Computer Science (), vol 9330. Springer, Cham. https://doi.org/10.1007/978-3-319-24306-1_38

[15] N. Kirthika and B. Sargunam, "YOLOv4 for Multi-class Artefact Detection in Endoscopic Images," 2021 3rd

International Conference on Signal Processing and Communication (ICPSC), 2021, pp. 73-77, doi: 10.1109/ICSPC51351.2021.9451761.

[16] F. Artunc and I. Oksuz, "An Ensemble Approach for Automatic Artefact Detection on Gastroendoscopy Images," 2021 6th International Conference on Computer Science and Engineering (UBMK), 2021, pp. 741-746, doi: 10.1109/UBMK52708.2021.9558919.