CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Report preparation needs meaningful frames which must be retrieved from the recorded video. It becomes tedious due to the presence of multiple artefacts in a single frame. On the other hand, if frames with artefacts are discarded, it reduces the video information content and thus affects video mosaicking quality. Hence detection is more helpful when an artefact restoration procedure is carried out.

Most researchers used an object detection algorithm to localize the artefacts in the endoscopic image. Object detection can be classified under visual recognition problems in CV. In the past decades, many researchers proposed novel techniques to precisely locate an object of targeted classes (ground truth) on the image and map each detected object to one of the classes specified. With the meteoric development of Deep Neural Networks (DNNs) (https://www.bmc.com) for object detection, the performance of the object detectors also greatly improved. Enablers of DNN performance include sophisticated algorithms, data availability, more computation power offered by GPUs and data augmentation techniques. In recent years GANs have created synthetic images very similar to original datasets images, which aids dataset expansion. It also boosts the algorithms performance as the DL algorithms rely on input data.

In the same way as artefact detection, artefact segmentation also plays a competent role. Perfect segmentation of the artefact aids an efficient restoration technique. Once an artefact is detected, the region must be segmented and restored. A restored image would aid better visualization the organ thereby reducing clinician procedure time, and also reduce faulty diagnosis results. Segmentation algorithms can be traditional or AI based algorithms.

2.2 METHODOLOGY OF REVIEW FOR ARTEFACT DETECTION

The review process started with searching algorithms in object detection, especially for health care applications. After reviewing articles from various health care fields, the

search narrowed to endoscopy. Numerous AI based applications in the field of endoscopy like polyp detection, cancer identification, polyp characterization, polyp classification, single and multi-class artefact detection, restoration of artefacts, diagnosis of helicobacter pylori and depth assessment of gastric cancer are studied. Multi-class artefact detection is found interesting due to limited researches and exciting challenges. At this stage, the search is restricted to deeply dive into the area of multi-class artefact detection and moved on for an extensive survey.

2.3 HISTORY OF OBJECT DETECTORS

Object detection refers to a CV task for identifying, localizing, drawing a bounding box and classifying the objects present in the bounding box. Object detection forms a base for other CV tasks like object tracking and image captioning. Every digital image and video contain an instance of semantic objects that belong to a specific class, such as, a bag, car, laptop or mug which has to be localized and classified. In the present decade, DL - based neural network algorithms are much preferred. This section reviews the state-of-the-art object detectors used in endoscopic artefact detection.

2.3.1 Single-Stage Object Detectors in Endoscopic Artefact Detection

- ***YOLO***

YOLO (J. Redmon, S. Divvala, 2016), divides images into regions, predicts bounding boxes, and simultaneously predicts probabilities, hence the name single stage detector. YOLO achieved attention as it achieved high accuracy while it could work in real-time. The network uses only one forward pass to make necessary predictions. The network proposes multiple bounding boxes, and class probabilities will be predicted for each box. NMS suppresses unnecessary bounding boxes and ensures the algorithm detects each object once. Finally, the network output will be a bounding box over a semantic object to be recognized. YOLO successors include YOLO9000 (Redmon & Farhadi, 2017), YOLOv3 (Redmon & Farhadi, 2018), YOLOv4 (Bochkovskiy et al., 2020),YOLOv5 and recently YOLOv8 is also brought into practice.

Sharib Ali and his team (Ali et al., 2021) trained variants of YOLOv3 and proposed YOLOv3-Spatial Pyramid Pooling (SPP). The trained network outperformed with best

results considering the performance parameter like mean Average Precision (mAP) and detection time. The network produced good Average Precision (AP) scores on detecting miscellaneous artefacts and bubbles, reaching a reasonable computational time of 88ms. YOLOv3 network results has been used as a baseline for their research (Y. yi Zhang & Xie, 2019). A network combining YOLOv3 and Mask R-CNN is proposed for artefact detection (Watanabe et al., 2019). In the model network proposed by the author, YOLOv3 is used to detect specific artefacts like blur and contrast as the network can maintain spatial relationships between objects and background. Many authors trained YOLOv3 for their research (Gao et al., 2019) (Huynh & Boutry, 2020). The results presented by authors based on the parameters like mAP, IoU and $Score_d$ are tabulated in Table 2.1. In the Table 2.1 $Score_d$ is calculated using Equation 2.1.

$$Score_d = 0.6\, mAP + 0.4\, IoU \tag{2.1}$$

The formula to calculate mAP and IoU is discussed in chapter 4.

Table 2.1 Performance of YOLO Variants in Endoscopic Artefact Detection

S.No.	Reference	Neural Network Architecture	Backbone	Performance Score
1.	(Ali et al., 2021)	YOLOv3	Darknet53	mAP= 0.351 IoU=0.242
		YOLOv3-SPP	Darknet53	mAP = 0.347 IoU=0.244
2.	(Gao et al., 2019)	YOLOv3 (threshold=0.1)	Darknet53	mAP=0.1750 IoU=0.2273
		YOLOv3 (threshold=0.25)	Darknet53	mAP=0.1668 IoU=0.2687
3.	(Huynh & Boutry, 2020)	YOLOv3	Darknet53	$score_d$=0.1702

- ***RetinaNet***

The RetinaNet architecture achieved state-of-the-art performance in 2017 International conference on CV (Lin et al., 2020). The architecture incorporates focal loss

to tackle the class imbalance problem. It contributes more to complex negative examples than easy examples, thus helping to improve the prediction accuracy. RetinaNet uses ResNet (K. He et al., 2016) and FPN (Lin et al., 2017) as its backbone. It also combines two task-specific networks for classification and bounding box regression.

RetinaNet with ResNet50 and FPN as backbone is adopted as one of the models for an ensembled architecture (Jadhav et al., 2020). Gao and his team (Gao et al., 2019) built an architecture based on RetinaNet for detection and instance segmentation. Sharib Ali and his team trained RetinaNet with Resnet50 backbone (Ali et al., 2021). A RetinaNet network with ResNet50 and ResNet101 as feature extractors is trained for artefact detection (Subramanian & Srivatsan, 2020). Both the models are used to predict outputs for the original and augmented image. The results are ensembled for the final predictions. The authors proved that RetinaNet with ResNet101 backbone has a higher mAP score. A RetinaNet based architecture (Vishnusai et al., 2020) extracts baseline results. The authors (Mohammad Azam Khan et al., 2020) optimized the RetinaNet network with weights pre-trained on the ImageNet dataset, applied various data augmentation techniques, employed hyperparameter tuning strategies to obtain the best detection score.

Maxime Kayser and his team proposed an ensemble method combining seven different RetinaNet architectures (Kayser et al., 2019) where every network varies in the backbone, hyperparameters, transfer learning strategies, training subsets and data augmentation technique used. The authors concluded that ensembled architecture along with incorporating other optimization techniques yielded the best score. A FPN is built on top of ResNet-152 as the backbone for RetinaNet (Ilkay Oksuz, et al.,, 2019). The authors modified variable parameters of focal loss and used them across all 100k anchors in each sampled image. The authors generated a massive dataset employing several data augmentation techniques. They preferred five-fold cross-validation. RetinaNet architecture is used as one of the models during the design of an ensembled architecture (Polat et al., 2020). Table 2.2 presents the performance scores obtained by various RetinaNet based object detection networks.

Table 2.2 Performance of RetinaNet Variants in Endoscopic Artefact Detection

S.No.	Reference	Neural Network Architecture	Backbone	Performance Score
1.	(Jadhav et al., 2020)	RetinaNet	ResNet50+FPN	mAP= 0.2607
2.	(X. W. Gao & Qian, 2019)	RetinaNet	ResNet101	$score_d$=0.2205
3.	(Ali et al., 2021)	RetinaNet	ResNet 50	mAP=0.347
4.	(Subramanian & Srivatsan, 2020)	RetinaNet	ResNet101 without TTA*	mAP=0.2151
			ResNet 50+ResNet101 with TTA	mAP=0.1537
5.	(Mohammad Azam Khan & Jaegul Choo, 2020)	RetinaNet	ResNet 101	mAP=0.2581 IoU=0.333 $score_d$=0.288
6.	(Kayser et al., 2019)	RetinaNet	ResNet50, ResNet101, ResNet152	mAP=0.3087 IoU=0.3997 $score_d$=0.3451
7.	(Ilkay Oksuz, James R. Clough & Andrew P. King, 2019)	RetinaNet	ResNet152	mAP=0.2719 IoU=0.3456

* - Test Time Augmentation

2.3.2 Two-Stage Object Detectors in Endoscopic Artefact Detection

- ***R-CNN***

R-CNN network (Girshick et al., 2014) is the first of its kind to use a CNN into an object detection network for higher performance. Later fast version of R-CNN called Fast R-CNN came into existence in the year 2015. Fast R-CNN extracts features for the entire image first and then sends them to the Region of Interest (RoI) pooling layer to extract the fixed size features, which are then sent to the classification layer and bounding box regressor. After a few months, an updated version called Faster R-CNN is developed. Later in the year 2015, an efficient Region Proposal Network (RPN) (Ren et al., 2015) which predicts region proposals with a vast range of scales and aspect ratio is deployed for usage. Mask R-CNN (K. He et al., 2020) is an extension of Faster R-CNN with a parallel branch, for instance segmentation. The author used Faster R-CNN with ResNet and FPN as the backbone for extracting features and replaced RoI pooling with RoI align to improve accuracy. Many researchers used networks belonging to the R-CNN family for endoscopic artefact detection.

A Faster R-CNN model is chosen and a right training strategy is applied to improve the network performance (Wang & Wang, 2019). Randomly initialized weights are used for classification and regression head, whereas for others, weights trained on Microsoft Common Object in Context (MSCOCO) dataset (https://cocodataset.org/) is preferred. To counter balance the class imbalance problem, patches are cut from images at the right scale and scaled up or down depending upon the object's size that is cut as a patch from the original image, which serves as a data augmentation technique. Detector performance is further boosted by replacing deformable convolution operation instead of regular convolution layers in FPN. A new scheme combining Mask R-CNN and YOLOv3 (Seiryo Watanabe et al., 2019) for detecting endoscopic artefacts is proposed. The authors reported that the Mask R-CNN network loses spatial relation between object and non-object regions. Thus, Mask R-CNN is preferred for artefacts with only clear and defined boundaries like specular reflection, saturation, bubbles, instruments and miscellaneous artefacts.

The author (P. Zhang et al., 2019) initially trained existing Mask R-CNN with images from the training set of segmentation task, where the instance masks are bounded by bounding box from the training set of detection task. The trained model will be used to predict masks for images from the training set of detection task, called soft-pixel level labels. The augmented dataset now comprises images from the original segmentation set with binary mask and training images from the detection task with soft pixel-level labels. The Mask R-CNN is retrained with an augmented dataset and is now called a Mask aided R-CNN. Later Mask aided R-CNN with various backbone architectures liket ResNet50, ResNet50+FPN and ResNet101+FPN are trained with the augmented dataset for building an ensembled architecture. The author reported that improving the network's performance with soft pixel-level labels is not much explored. Existing Fast R-CNN network with ResNet101 as the backbone is trained for the detection task (X. W. Gao & Qian, 2019). A Faster R-CNN based on ResNet (He, K., et al., 2015) and ResNeXt (Xie, S., et al., 2016) modules is trained for artefact detection (Polat et al., 2020). The author tuned various hyperparameters to produce a state of the art results. Table 2.3 details the results of various two stage object detection networks and its performance on endoscopic artefact detection.

Table 2.3 Performance of R-CNN Variants in Endoscopic Artefact Detection

S.No.	Reference	Neural Network Architecture	Backbone	Performance Score
1.	(Wang & Wang, 2019)	Faster R-CNN	ResNet50+ FPN	mAP= 0.2621 IoU=0.3205 $Score_d$=0.2855
2.	(Watanabe et al., 2019)	Mask R-CNN+ YOLOv3	ResNet101 + FPN	mAP= 0.2901 IoU=0.318 $Score_d$=0.3013
3.	(P. Zhang et al., 2019)	Ensembled Mask Aided R-CNN	ResNet50, ResNet101 + FPN	mAP= 0.3117 IoU=0.4051 $Score_d$=0.361
4.	(X. W. Gao & Qian, 2019)	Fast R-CNN	ResNet 101	mAP=0.2416 IoU=0.3482 $score_d$= 0.2842
5.	(Vishnusai et al., 2020)	Faster R-CNN	ResNeXt 101+FPN	$Score_d$=0.2319

2.3.3 Multi-Stage Object Detector in Endoscopic Artefact Detection

Multi-stage detectors aim at achieving better accuracy than one-stage and two-stage detectors. Cascaded R-CNN (Cai & Vasconcelos, 2018), an extended version of faster R-CNN, is said to overcome the problem of overfitting during training and inference time mismatch between IoUs. IoU threshold is said to have a greater impact to classify positive and negative samples. The idea behind the design is that, various IoUs are set to train the model. Basic R-CNN based models are cascaded where the output of the previous detection model is set as input to the next detection model. The IoU is said to keep increasing as the stage progress.

A basic cascaded R-CNN with L1 loss function is chosen for artefact detection (Ning et al., 2019). The authors (Y. yi Zhang & Xie, 2019) proposed a multi-stage cascaded R-CNN combined with FPN. The authors used a phased approach to gradually increase the IoU threshold during training. Initially, the model is pretrained with images from the MSCOCO dataset and later trained using the EAD 2019 dataset (Ali et al., 2019). Various data augmentation techniques are involved. To improve the performance Chain method is incorporated where binary classifiers use predictions of all previous stages to produce the result of the current stage.

An improved version of cascaded R-CNN structure by adding ResNet101 as the backbone along with the FPN module is proposed (S. Yang & Cheng, 2019). The model has two main sub-modules: multi-scale feature extractor and multi-stage object detector. The former module extracts the best features to improve the detection rate. The author used t-distributed Stochastic Neighbour Embedding (t-SNE) to visualize the data distribution of the dataset. Outliers are removed based on the observations, data augmentation techniques for few samples concentrating on a few artefacts, namely saturation and blur, are employed. This proposed model reported a good balance in performance between mAP and IoU.

An ensemble model is proposed in which cascaded R-CNN is one among the three models used (Polat et al., 2020). The prediction of all three models, namely Faster R-CNN, Cascaded R-CNN and RetinaNet, are fed into class agnostic NMS to remove redundant bounding boxes. It is followed by an ensemble, after which false-positive elimination is done to improve the performance. This method proposed by the author has the top performance of 20.31% for mAP and 32.85% for IoU. A cascaded R-CNN with FPN and ResNeXt101 networks to extract features, including Deformable Convolutions (DCN) is experimented (Hung et al., 2020). A resampling mechanism is used to reduce overfitting. DCN is added to the backbone at stage 3 to stage 5 to differentiate background from desired objects. This helps to improve the performance. A cascaded R-CNN network with ResNeXt and FPN network from MM Detection Toolbox is used for artefact detection (Hu & Guo, 2020). Soft-NMS is used to avoid over-detecting objects in the image. Images are resized to 1024 x 1024 for effectively detecting small objects. A cascaded R-CNN is a base model with ResNet101 backbone network trained on ImageNet dataset along with FPN for artefact detection (Yu & Guo, 2020). Various training strategies are used for data augmentation, modified loss function, cosine decay learning rate schedule and box ensemble techniques. Soft-NMS is replaced instead of regular NMS operation. The author quoted that the box ensemble performs poorly as it causes lower mAP. Table 2.4 briefs the results obtained by authors based on the Cascaded R-CNN network.

Table 2.4 Performance of Cascaded R-CNN Variants in Endoscopic Artefact Detection

S.No.	Reference	Neural Network Architecture	Backbone	Performance Score
1.	(Ning et al., 2019)	Cascaded R-CNN	ResNet101	IoU=0.1222 mAP=0.3068 $Score_d$=0.2330
2.	(Y. yi Zhang & Xie, 2019)	Cascaded R-CNN	FPN	$Score_d$=0.3429
3.	(Polat et al., 2020)	Cascaded R-CNN	ResNet+ FPN	IoU=0.3221 mAP=0.2996 $score_d$=0.3086
4.	(Polat et al., 2020)	Ensemble of Faster R-CNN, Cascaded R-CNN and RetinaNet	ResNet50+FPN	IoU=0.4591 mAP=0.4571
5.	(Hung et al., 2020)	Cascaded R-CNN	ResNeXt101+FPN	$Score_d$=0.2366
6.	(H. Hu & Guo, 2020)	Cascaded R-CNN	ResNeXt+FPN	$Score_d$=0.2202
7.	(Yu & Guo, 2020)	Cascaded R-CNN	ResNet101+FPN	$Score_d$=0.2036

2.3.4 Anchor Free Detectors

Traditional DL - based architectures heavily rely upon anchors for predicting semantic objects in an image. These anchors have various scales and aspect ratios based on the object to be detected in an image. It is said that the speed and accuracy of the detector are based on the anchors, where fewer the anchors faster the detectors, but it may reduce accuracy. At the same time, it involves a lot of hyper-parameters which directly affects the IoU score. Thus, anchor free detectors took advantage over the existing object detection algorithms.

Anchor free object detectors work on the principle of key-point detection. It generates a heatmap using CNN and relies on NMS to suppress unnecessary bounding boxes. It is said that anchor free object detectors find detecting a dense and overlapping object difficult. Present-day anchor free object detectors include CornerNet (Law & Deng, 2020), CenterNet (Duan et al., 2019), Fully Convolutional One-Stage object detection (FCOS) (Tian et al., 2019), mask attention anchor free detection (H. Yang et al., 2020) and Paddle Anchor Free Network (PAFNet) (Xin et al., 2021). Detection of endoscopic artefacts using anchor free object detectors is at its budding stage, and there is a large scope for researchers to explore.

2.4 METHODOLOGY OF REVIEW FOR ARTEFACT SEGMENTATION

The review of image segmentation started with seeking literature articles particularly in the health care field. A parallel line for endoscopic artefact detection, the artefact segmentation is chosen to develop an end to end framework for efficient artefact restoration. Hence, a deep dive into artefact segmentation started with articles and researches related to single and multiple artefacts. This section on literature review analyses recently reported articles on the segmentation of single and multiple artefacts.

2.5 HISTORY OF SEGMENTATION ALGORITHM

U-Net (Ronneberger et al., 2015), is an expanded version of FCN for biomedical image segmentation. Based on an FCN, the network only has convolutional layers, reducing the number of parameters and supporting any size images. The U-Net is trained from beginning to end and has shown to be a very successful method for a variety of applications where the size of the output is similar to the input size. The network design consists of two paths: an expansive track (decoder) to enable exact localization and the contracting path called as the encoder to extract features and record the context in the image. This gives the network layout a U-shaped appearance. The contracting path consists of two 3*3 convolutional layers, each followed by Rectified Linear Unit (ReLU) (Agarap, 2018), an activation function and a max-pooling layer. The max-pooling layer holds the size of 2*2 with a stride of two for down sampling. During down sampling the feature channels are doubled for every step.

Two 3*3 convolutions are used to concatenate the expanding and contraction path. To reduce the number of channels by two in the expansive path every block contains a 2*2 up-convolution and two 3*3 down convolutions. Each of them followed by ReLU. The network's last layer is a 1*1 convolution layer to map the feature map size from 64 to the required number of classes. To compensate the lost information the U-Net uses skip connections.

A team of researchers proposed the residual network names ResNet. The network through the “identity shortcut connection” enables one or more layers to be skipped by removing activation from one layer and transferring it to another layer. The network finds a unique solution to the issue of vanishing gradient. This network has different variants, which is based on number of layers 34, 50,101 and 152 layers. Later during the year 2017, dense convolutional network called DenseNet (G. Huang, et al., 2016) is proposed. The network

uses tensor concatenations, where the network grows denser with each skip connection. In the DenseNet, all the dense blocks are connected directly following a feed forward style. All the feature maps of previous layers are concatenated and passed to the next block. This architecture is proven for easy training and deep supervision.

In the same year 2017, ResNeXt, a variant of ResNet is proposed. The network follows a technique called split, transform and merge. This network proves to be more effective in terms of accuracy. Later in the year 2018, a Squeeze and Extraction (SE) block is introduced (J. Hu et al., 2017). This network can acquire the global weighting over all of the pixels in a channel. It also allows fast transferring of information between pixels. In 2019, a network called EfficientNet is proposed (Tan & Le, 2019). This network is named for its lightweight CNN based on Automated Machine Learning (AutoML) (X. He et al., 2019). This network has a different scaling method that will systematically scale up the width, depth and the resolution. A baseline version called EfficientNetB0 is developed. Later the network is scaled up with different version from B0 – B7. These networks are known to achieve best accuracy and computational efficiency.

Most of the image segmentation research related to the field of bio-medical utilizes U-Net architecture with popular backbone for efficient segmentation. The following section portrays different networks used for artefact segmentation especially the artefacts in endoscopic images.

2.6 SEGMENTATION MODEL FOR SNGLE ARTEFACT SEGMENTATION

A reliable and real-time surgical instrument segmentation algorithm required for endoscopic vision is proposed (Garcia-Peraza-Herrera et al., 2017; Qin et al., 2019). The method combines CNN predictions and kinematic pose information. ToolNet-C is trained with many unlabelled images, while the pixel-wise segmentation module is trained with a small number of labelled images. The kinematic pose calculates the projection of the instrument's body onto the endoscopic image. The kinematic pose is estimated and refined by a particle filter. The refined final pose determines the accurate mask, which is considered as the final segmented output.

A surgical tool segmentation algorithm is proposed by Kamrul Hasan and his team (Hasan et al., 2021). Using restoration algorithms, the authors demonstrated how to segment and remove surgical instruments from video. A U-Net+ architecture with a pre-trained Visual Geometry Group (VGG) encoder and decoder is used. Instead of transposed convolution, the decoder operation is modified with up-sampling based on nearest-neighbour interpolation. The tool removal algorithm employs the restoration technique, in which the instrument mask is filled with the tissue beneath it.

A method for removing specular reflections in each Red Green Blue (RGB) channel of an image using a simple thresholding technique is proposed (Lim, 2020). To ensure complete local segmentation, the author chose dilation. Image restoration is used to replace areas with reflections with background pixels that do not have reflections. The author created a visually enhanced image by incorporating advanced techniques such as, histogram shift, histogram equalisation, and gamma-correction. The author conducted their research using the Iparkmall Clinic dataset.

2.7 SEGMENTATION MODEL FOR MULTIPLE ARTEFACT SEGMENTATION

A Deeplabv3+ based network with two different backbones, ResNet101 and MobileNet is investigated for artefact segmentation (S. Yang & Cheng, 2019). After the addition of backbone network five parallel convolutional layers are added for feature extraction. The five layers include 1*1 convolutional layer and 3*3 dilated convolutional layers. The three dilated layers are designed with different rations as 6, 12, and 18 respectively. The last one layer is nothing but a pooling layer. In order to achieve region segmentation all the feature maps extracted are merged and up-sampled. The segmentation models are trained with single and merged backbones. The authors confirmed that merging the backbones improved algorithm performance with an overlap score of 0.6612.

U-Net architecture with various backbones VGG-11, VGG-16, PSPNet with residual structure, PSPNet with ResNet-34 and DeepLabV3 with residual structures are investigated for artefact segmentation (Y. yi Zhang & Xie, 2019). Out of all the chosen networks the authors proved that PSPNet with ResNet-34 architecture performed well across segmentation of all the artefacts. The authors relied on data augmentation techniques like resizing the training set images in scales like 0.7, 0.9, 1.0, 1.2 and 1.5 times that of the original image.

Later the scaled images are randomly cropped, then flipped horizontally and vertically and the images are rotated at various angles. Hence the dataset for training the segmentation algorithm expanded manifold. Few initializations like training with batch size of 8, weight decay of 0.01 and cyclical learning rate adjustment resulted with overlap score of 0.59 and F2 score of 0.62.

Deep Layer Aggregation structures (DLA) (Ning et al., 2019) for efficient artefact segmentation is proposed. This network merges features extracted iteratively and hierarchically. This method is said to improve the segmentation accuracy. The authors used a weighted multi-class dice loss as the "segmentation loss" to counterbalance the class imbalance problem. The performance of DLA is compared with the performance of U-Net based architectures.

A U-Net-based architecture is proposed (S. Yang & Cochran, 2019). CNN is used to extract spatial features, and transposed convolution is used as a decoder, with ResNet50 pre-trained on ImageNet as the backbone. The authors used a poly learning rate policy during the training process to improve segmentation performance and obtained an F2 score of 0.56.

A U-Net based Architecture for the purpose of multiple artefact segmentation is proposed (Mohammad Azam Khan & Jaegul Choo, 2020). This simple network is trained vigorously to achieve an overlap and F2 score of 0.43. OxEndoNet (Gridach & Voiculescu, 2020) uses Pyramid Dilated Modules (PDM) for its function. The network is made up of multiple stacked parallel dilated convolutions. The network is built with 3*3 convolutions with different dilation rates from 1 to 4. ReLU is used as an activation function. Combining dilated convolutions using different rates is said to improve algorithm performance. OxEndoNet is made up of many pyramid architectures.

A U-Net-based network with SE-ResNeXt50 as the backbone is explored (Hung et al., 2020). The authors chose the network because it addresses the problem of class imbalance and can retrieve hidden fragments in images. The binary cross entropy (BCE) loss and the dice loss are employed. Pre-trained weights are used to improve performance and decrease training time, yielding a segmentation score of 0.5700.

A DeeplabV3 network with Xception backbone and U-Net with ResNet50 backbone is designed (Subramanian & Srivatsan, 2020). An efficient and accurate scene text detector is used to detect the text in the endoscopic image, which is classified as a miscellaneous artefact in the dataset. Train and test time augmentation is also employed. For improved performance, an ensemble architecture combining both of the architectures mentioned detailed in the section is used in conjunction with a pixel-wise voting technique, yielding a best segmentation score of 0.4966.

The U-Net based network is trained using a combination of backbone and augmented data (Vishnusai et al., 2020). U-Net with ResNeXt50 backbone performed well in all experiments. The authors used various architectures from the segmentation model framework. The segmentation score produced by the trained architecture is 0.5187.

A U-Net and U-Net++ based architectures (Huynh & Boutry, 2020) is proposed by Huynh and his team. EfficientNetB1 is used as backbone which balances resolution, network depth, and width. It extracts feature maps at five scales and feeds into the decoder stage. Several data augmentation techniques and ImageNet pre-trained weights are used to reduce training time and avoid overfitting. To improve the results, test time augmentation is also used.

A multi-plateau ensemble of FPN with EfficientNet as backbone to extract features is proposed for artefact segmentation (Jadhav et al., 2020). The EfficientnetB3, EfficientnetB4 and EffcientnetB5 optimizers like Ranger and Over9000 with various loss functions such as, dice, BCE + dice, BCE, BCE + dice + Jaccard mapping to each optimizer is trained and tested for performance. Total of 24 different models are trained for final ensemble. The model that achieves a dice score greater than 0.47 is given priority.

A DeeplabV3+ based network (Guo et al., 2020) is proposed for artefact segmentation. The network holds SE-ResNExt-50 as backbone. A global pooling layer is replaced by 3*3 convolution layer. This is expected to improve segmentation of small objects. Few progressive modifications improved the segmentation score.

The Table 2.5 compares the performance of all the available multiple artefact segmentation algorithm. The performance metric $Score_s$ is calculated using the Equation 2.2.

$$Score_s = 0.75 * [0.5 * (Dice\ Similarity\ Coefficient + Jaccard\ Score)] + 0.25 * F2 \quad (2.2)$$

Table 2.5 Summary of Performance of Artefact Segmentation Algorithms

S.No.	References	Network architecture	Performance score		
			Jaccard Score	F2 Score	$Score_s$
1.	(Suhui, Y., & Guanju, C., 2019)	DeepLab v3 with ResNet101 Backbone	0.6288	0.6795	0.6414
		Ensemble of DeepLab v3 with ResNet101 & MobileNet backbone	0.6592	0.6937	0.6568
		Ensemble + post processing	0.6612	0.6964	0.6700
2.	(P. Zhang et al., 2019)	Ensemble Mask Aided R-CNN with ResNet101 backbone	0.5397	0.5701	0.5594
3.	(Ning et al., 2019)	DLA-60(crf)	0.5206	0.5661	0.5320
		DLA-60	0.4352	0.4784	0.4460
4.	(S. Yang & Cochran, 2019)	U Net	0.36	0.48	0.42
		DeepLabv3+	0.6416	0.6779	0.55
		U-Net-D	0.39	0.44	0.41
5.	(Gridach & Voiculescu, 2020)	OxEndoNet	0.4901	0.5107	0.5194
6.	(Hung et al., 2020)	U-Net+SE ResNext50 backbone	-	-	0.5700
7.	(Subramanian & Srivatsan, 2020).	Ensemble of DeepLab + U-Net + East text Detector	Dice Score 0.42946		0.4966
		DeepLab	0.39998		0.4574
8.	(Vishnusai et al., 2020).	U-Net+ResNext50 backbone	-	-	0.5187
9.	(Yun, B. G., et al., 2020)	Net 1- DeepLabv3 + SEResNeXt 50	-	-	0.48
		Net 2- Net 1 with 3*3 convolution instead of global pooling	-	-	0.59
		Net 3- Net 2 with squeeze and extraction module	-	-	0.52
		Net 3 with test time augmentation	-	-	0.5922
10.	(Huýnh, L. D., & Nicolas, B., 2020)	U-Net++ with Efficient Net B1 backbone	-	-	0.5913
		U-Net++ with Efficient-Net B2 backbone	-	-	0.5397
11.	(Jadhav et al., 2020)	Multi-Plateau ensemble of FPN and Efficient-Net B3	Dice Score- 0.4917		

2.8 RESEARCH GAPS IDENTIFIED

The research gap identified are listed below,

a. The EAD dataset is the only available public dataset for artefact detection and segmentation. The dataset suffers from class imbalance problem and repeated frames.
b. The detection and segmentation of multiple artefacts through deep learning algorithms are at its novice state due to limited dataset and research.
c. The algorithms can be integrated with endoscopic imaging pipeline only if the inference time is reduced. It is not addressed by many of the researchers.
d. An efficient system that combines artefact detection, segmentation, restoration and disease identification could be very effective for real time implementation, which very few researchers have implemented. The fully combined pipeline is not reported in the literature.

2.9 CHAPTER SUMMARY

The literature articles recited in this thesis stick to a subset of AI called DL in endoscopy. The need for high-performance artefact detection systems in the endoscopic imaging pipeline is becoming essential daily. Achieving very high accuracy, greater than 95% has become the ultimate goal for almost all researchers. Researchers follow various strategies include training from scratch, transfer learning, extracting features rich in spatial information, modifying existing architecture, proposing a new architecture, varying backbones, proposing ensemble method, incorporating pre and post-processing techniques, train and test time augmentation. The profound potential of such AI-assisted technologies in the field of endoscopic imaging can transfer the clinical practice efficiency and accuracy to uncover relevant information from data to help the endoscopist for better clinical decision making in the future. Throughout this literature study, it is observed that just a couple of researchers are working towards endoscopic artefact detection.

Most of the present articles concentrate only on accuracy. But inference time is also considered to be an equally important parameter. There are a lot of arenas open for autonomous surgical bots for minimally invasive surgery. Pertaining to segmentation most

researchers used DL techniques to segment either single or multiple artefacts. The performance of traditional segmentation algorithms in segmenting a variety of endoscopic artefacts is never considered. To improve performance, these traditional algorithms can be combined with modern algorithms. Most researchers focused on evaluation metrics such as, dice score and Jaccard index. Other metrics such as, sensitivity, specificity and accuracy are also essential. Furthermore, most DL algorithm researchers prefer U-Net-based architectures and DeepLabV3 whereas other segmentation architectures such as, FPN and Link-Net are not much used.