CHAPTER 3
DATASET

3.1 IMPORTANCE OF THE DATASET

AI partnered applications are at their peak of popularity today. The lifeline of many real-time AI-based applications is in the hands of well-annotated datasets and training hardware. DL algorithms do not need manual feature extraction like ML algorithms; instead, they learn from the labelled data. Thus well annotated dataset is essential. DL algorithms pine for massive datasets to train themselves.

Crafting a new dataset involves various processes such as, data (image) collection, data cleaning, formulating annotation protocols, selecting the best tool, label output format, listing labels (meaningful objects). Labelling dataset consumes almost 60% of the research time. Thus most researchers use public datasets, for applications involving natural images lot of public datasets are available, such as, MSCOCO, ImageNet (https://www.image-net.org) and Canadian Institute For Advanced Research (CIFAR)10 (https://www.cs.toronto.edu). Application-specific datasets like Modified National Institute of Standards and Technology (MNIST) (http://yann.lecun.com) for handwritten digits recognition, Internet Movie Database (IMDb) reviews dataset (https://www.imdb.com) for Natural Language Processing (NLP) and Date Fruit dataset (https://www.kaggle.com) for classification of date fruits based on genetic varieties using Image Analysis are also available.

On the other hand, many researchers who work in a constricted area of research tend to create a new dataset. Similarly, hardware requirements for training DL-based algorithms are prohibitive. GPUs are required to do the job. Due to higher GPU cost, cloud-based training (https://cloud.google.com) is gaining popularity.

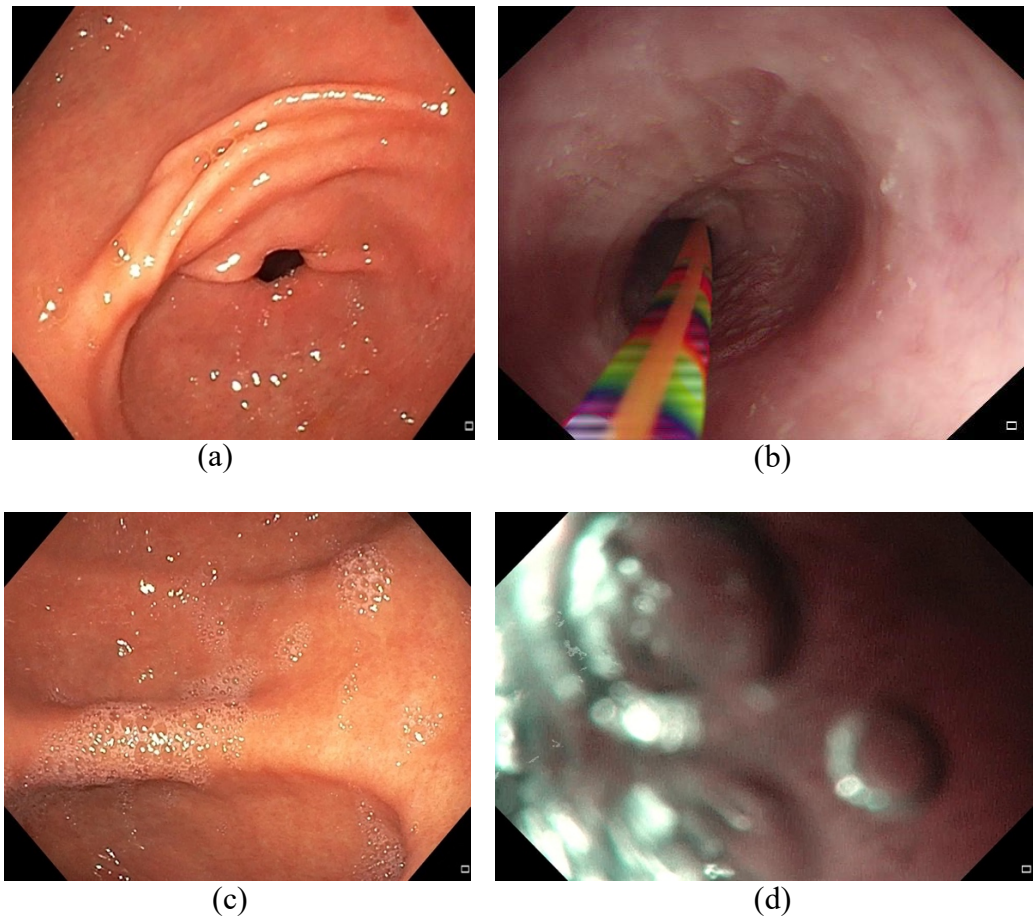
3.2 PUBLIC DATASET

To study on particular applications like artefact detection or segmentation in endoscopic images, several datasets are publicly available. All the available dataset covers only a single artefact, namely the Kvasir-Instrument dataset (https://datasets.simula.no) for endoscopic instruments, CVC-ClinicSpec for specular reflection detection and Cholec80

dataset (http://camma.u-strasbg.fr/datasets) for instrument detection and tracking. So researchers in the last decade concentrated on detecting a single artefact where a few researchers even restored the artefacts. Such datasets will not be sufficient to restore affected images for a qualitative analysis of the entire endoscopic video. Artefacts like blur and specular highlights corrupt every image produced during the procedure. The importance lies when several artefacts are found to be present in the same image. Hence detection and segmentation of one artefact do not give a consequential research outcome. The need to deal with multiple artefacts arises then. In 2019 and 2020, EAD datasets came into existence. The dataset is dedicated to multi-class artefact detection, segmentation and generalization. The dataset motivated many researchers to take the challenge of dealing with various artefacts present in the image in a single shot.

3.2.1 EAD2019

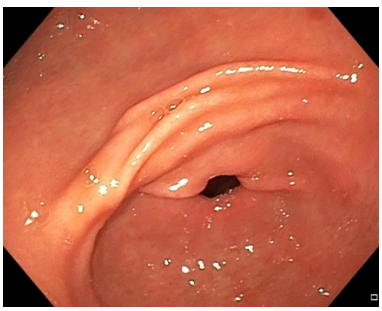
EAD2019 is the first comprehensive dataset for multi-class artefact detection, segmentation and generalization. This dataset holds images of multiple organs obtained from multiple patient. The dataset holds images taken using multiple modalities. The organ under study includes the oesophagus, stomach, colon, liver and bladder. The images in the dataset are imaged using standard endoscopes. It comprises images taken in various imaging modalities such as, white light, NBI, and AFI. The dataset embraces images containing seven major artefacts such as, specular highlights, saturation, contrast, blur, bubbles, instruments and miscellaneous artefacts. For artefact detection, 2147 images are available in the training dataset and 475 images are available for artefact segmentation. Binary masks for five different artefacts, including saturation, bubbles, specular reflection, instrument and miscellaneous artefacts, are available for artefact segmentation. Sample images with ground truth annotations from the EAD2019 detection dataset are shown in the Figure 3.1 (a) – (d). Figure 3.2 (a) and 3.3 (a) portrays images from the EAD2019 dataset. The corresponding binary mask for each of the artefact portrayed in Figure 3.2 (a) and Figure 3.3 (a) is given in the Figure 3.2 (b) – (f) and in Figure 3.3 (b) – (f).



(a) (b)

(c) (d)

Figure 3.1 (a) - (d) Sample Images from EAD2019 Detection Dataset



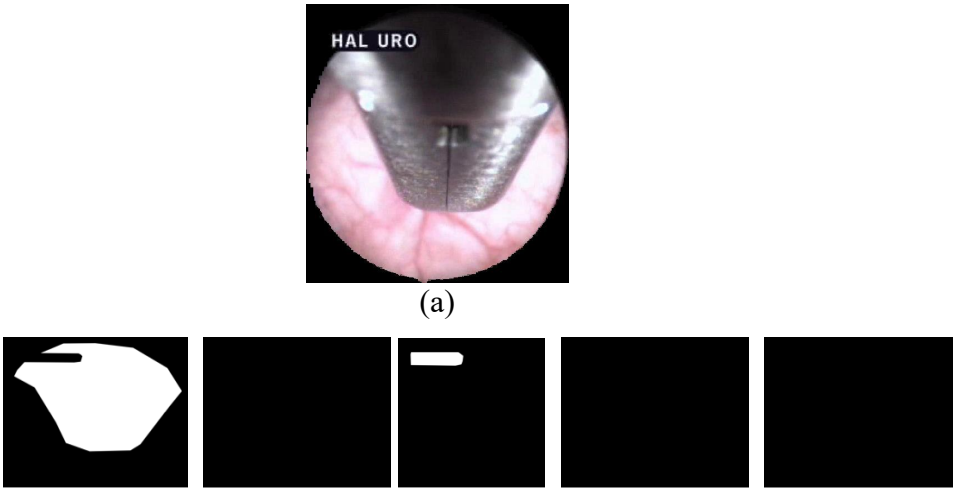
(a)

(b) (c) (d) (e) (f)

Figure 3.2 (a) A Sample Image from EAD2019 Segmentation Dataset

Figure 3.2 (b) - (f) Corresponding Binary Mask for Each Artefact

In the Figures 3.2 and 3.3 (b) – (f), (b) holds the mask that corresponds to the instrument artefact. (c) holds the mask of specular reflections, (d) belongs to miscellaneous artefact, (e) belongs to saturation and (f) belongs to bubbles. The masks for artefacts that not present in the Figures 3.2 and 3.3 (a) are intentionally left blank.



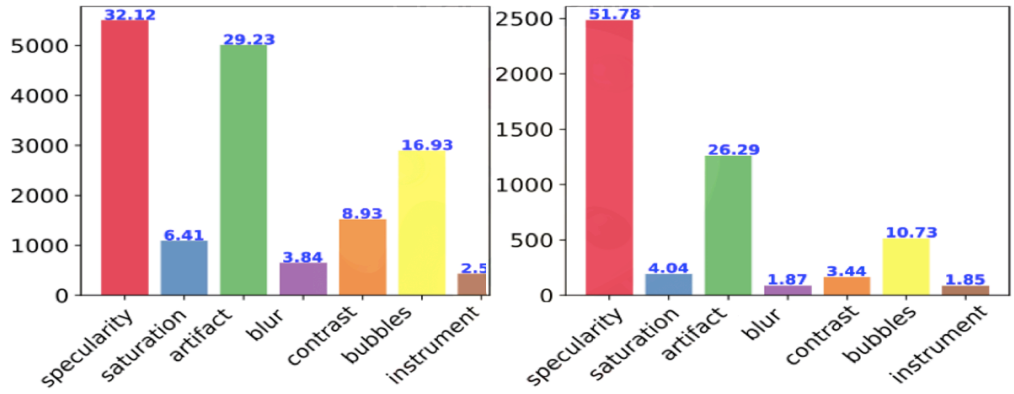
(a)

(b) (c) (d) (e) (f)

Figure 3.3 (a) A Sample Image from EAD2019 Segmentation Dataset

Figure 3.3 (b)-(f) Corresponding Binary Mask for Each Artefact

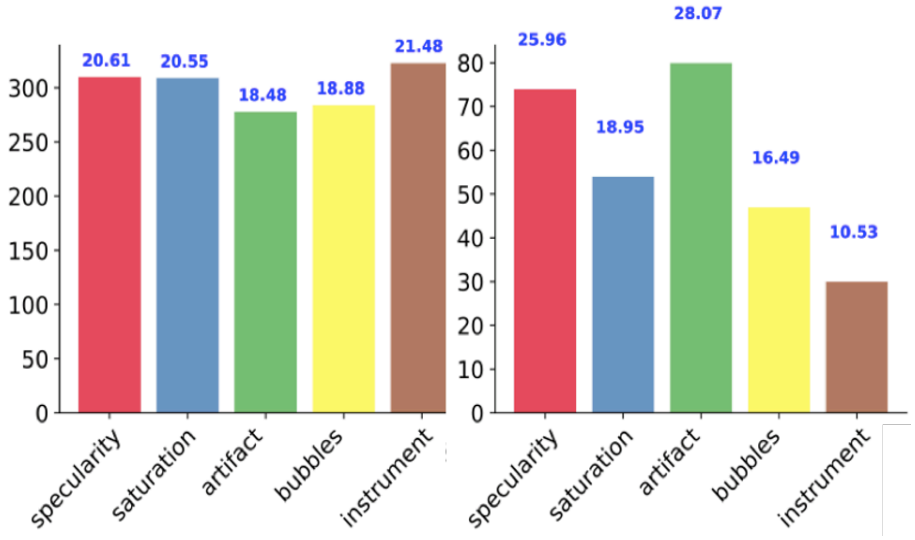
The class-wise allocation of artefacts in training and test-set for detection tasks is represented graphically in the Figure 3.4 (a) and (b). Artefact specular reflection and miscellaneous artefacts dominate the distribution. On the other hand, artefacts such as, instruments and blur are found sparsely distributed in the dataset. The same scenario is also present in the class-wise distribution of artefact in the test set.



(a) (b)

Figure 3.4 (a) and (b) Class–wise Distribution of Artefact in EAD2019 Training and Test Set for Artefact Detection (Ali et al., 2019)

Figure 3.5 (a) and (b) depicts the class-wise allocation of artefacts in training and test set for segmentation purposes. The distribution is found almost balanced among all five classes in both the training and test set.



(a) (b)

Figure 3.5 (a) and (b) Train and Test Data Distribution of Segmentation Dataset per Artefact (Ali et al., 2019)

3.2.2 EAD2020

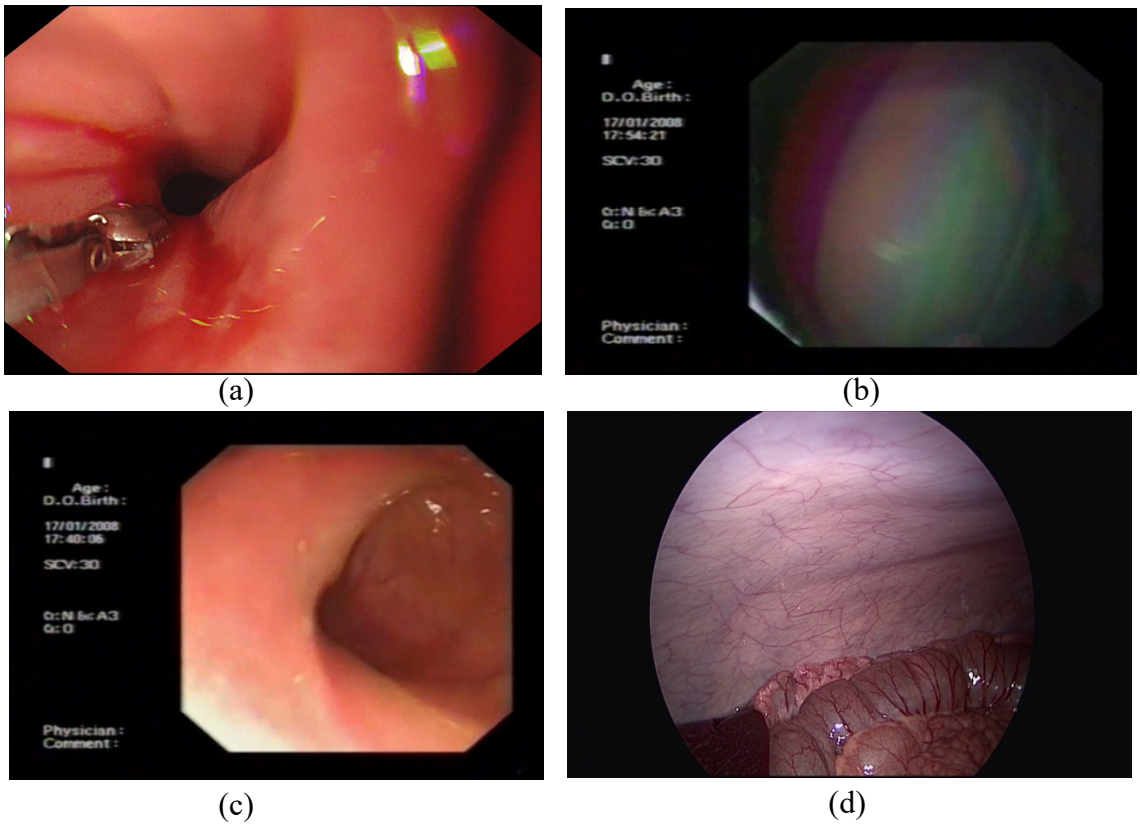
EAD2020 is an extension of the EAD2019 dataset, in which one more artefact called blood is added to the dataset. It encompasses 2531 images. Sample images from the EAD202 dataset is displayed in the Figure 3.6 (a) – (d). All the artefacts are suggested by the Clinicians.

The images are gathered from six different medical facilities across the world. The dataset includes images of multiple organs, including oesophagus, stomach, liver, colon, and bladder. It holds images of multiple tissue (colonoscopy, cystoscopy, gastro-oesophageal, and white light) and contains images taken using multiple modality such as, AFI, white light and NBI. The images are taken with regular endoscopes. The photos stored in the collection do not contain any patient data.

Senior clinicians labelled the images in the initial stages later on by the proficient post-doctoral fellows. Clinicians validated the images at the end. Bounding box annotations

for artefact detection is done with in house custom tools. For semantic segmentation, the dataset holds 643 images with binary masks for corresponding artefacts. The dataset also holds images for generalization tasks.

Binary masks for endoscopic artefacts such as, specular highlights, saturation, bubbles, instrument and miscellaneous artefacts are present in the segmentation dataset. According to many researchers the dataset has a class imbalance and the reason is due to the irregular distribution of diverse artefacts in a single image. Specular reflections are scattered across the image in a random fashion and similarly miscellaneous artefacts are found in most of the images next to specular highlights. Both of the artefacts are randomly distributed in a single frame varying from small to large. All possible artefacts are bound using a separate bounding box to avoid overestimation of the bounding box. Hence the number of bounding boxes per artefact in all images rapidly increases, especially for specular reflections. Hence class imbalance problem becomes unavoidable.

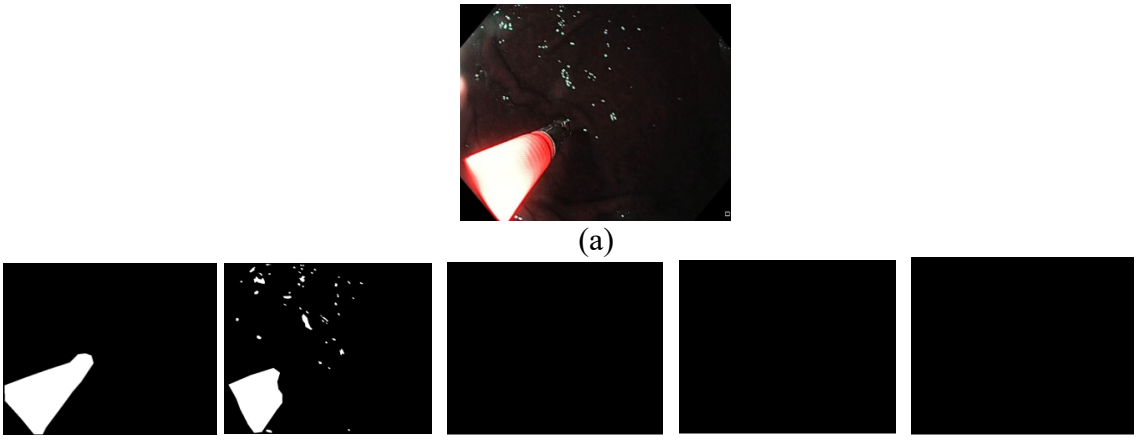


(a) (b)

(c) (d)

Figure 3.6 (a)-(d) Sample Images from EAD2020 Detection Dataset

Figure 3.7 (a) shows an image given in the EAD2020 dataset for segmentation. The image is affected by several artefacts. Especially specular reflection and instrument overlap. Hence one single binary mask cannot to be used. For ease separate binary masks are given for each artefact as displayed in the Figure 3.7 (b) – (f).



(a)

(b) (c) (d) (e) (f)

Figure 3.7 (a) A Sample Image from EAD2020 Segmentation Dataset

Figure 3.7 (b)-(f) Corresponding Binary Mask for Each Artefact

In the Figure 3.7 (b) – (f), (b) holds the mask that corresponds to the instrument artefact. (c) holds the mask of specular reflections, (d) belongs to miscellaneous artefact, (e) belongs to saturation and (f) belongs to bubbles. The masks for artefacts that not present in the Figure 3.7 (a) is intentionally left blank.

3.3 CURATION OF THE CUSTOM DATASET

Endoscopic images of patients belonging to western countries are found in the dataset. There aren't many images with artefacts such as, blur, saturation and instrument. Thus, the images of Indian patients are collected. In order to balance the dataset, a higher importance is given to images with saturation, blur and instrument.

The endoscopic video recorded during the procedure and a few shots acquired for report generation are pooled together. Due care is taken not to have patient information in any videos or frames taken for dataset preparation. The video sequence is split into individual frames. Frames are chosen from it based on the following criteria:

- More images affected by the artefact blur, instruments and saturation.
- Fewer number of images with artefact specular reflections.

- Scarcely frames are selected from a sequence to show variations and avoid repetitive frames.

Thus 2400 images are consequently gathered. Each of the images are analysed to identify the possible artefacts. The identified artefacts are specular highlights, saturation, blur, bubbles, blood, contrast, instruments and miscellaneous artefacts. Images are initially annotated before a senior clinician to acquire knowledge. The scholar does later annotations and the clinician validated the annotations in the ratio of 1:10 in terms of number of images. A selection of the custom dataset's images are shown in Figure 3.8 (a) through (d).

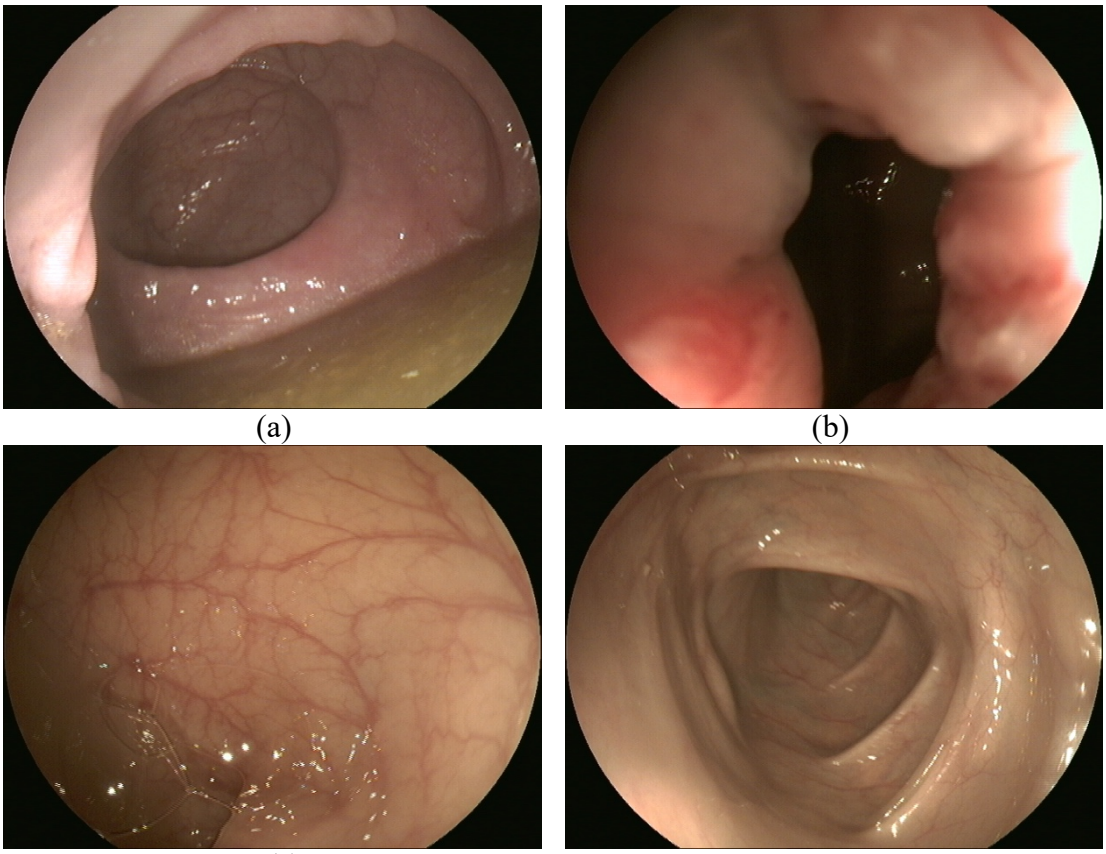


Figure 3.8 (a)-(d) Sample Images from Custom Dataset

3.3.1 Annotation protocols

The custom dataset is annotated using the same procedures as the EAD2019 dataset. The endoscopic artefact detector is also trained using images from EAD datasets. Uniform

annotation protocols are used to maintain homogeneity across both public and custom datasets. The protocols are as follows:

- Multiple boxes are used to annotate if the region belongs to more than one artefact (class).
- For artefact like specular reflection, instead of using one single bounding box, many small bounding boxes are used to avoid over-estimation of the bounding box.
- Distinctive artefacts are identified, and bounding boxes are drawn.
- Each of the identified artefacts is ascertained to be general across endoscopy datasets.

Images from a custom dataset annotated for eight frequently occurring artefacts are shown in Figure 3.9 (a)–(d). A wider region is covered by artefacts such as, instruments, saturation, blur, blood, bubbles, contrast, and a few more miscellaneous artefacts. Specular reflections and a few other miscellaneous artefacts only cover a small portion of the image. Most of the specular reflections are marked with separate bounding boxes for precise delineation. A single bounding box is scarcely used when the specular reflections are located in a sequential fashion. In Figure 3.9 (a), artefact such as, specular reflection, contrast and blur are annotated. In Figure 3.9 (b) artefacts such as, bubbles, specular reflections and instrument are identified and annotated. Figure 3. 9 (c) is affected by miscellaneous artefacts, blur, bubbles and specular reflections. Similarly, Figure 3.9 (d) is affected by specular reflections, bubbles and saturation. From the sample images it is evident that, the features of specular reflections and miscellaneous artefacts are very similar. Number of bounding boxes used to annotate tiny specular reflections are more. Hence it is a trivial task to balance the dataset where apparently it is true that number for bounding box annotations for specular reflections will be always high. Additionally, it is clear from the images in Figure 3.9 (a) - (d) that almost every frame apparently contains multiple artefacts.

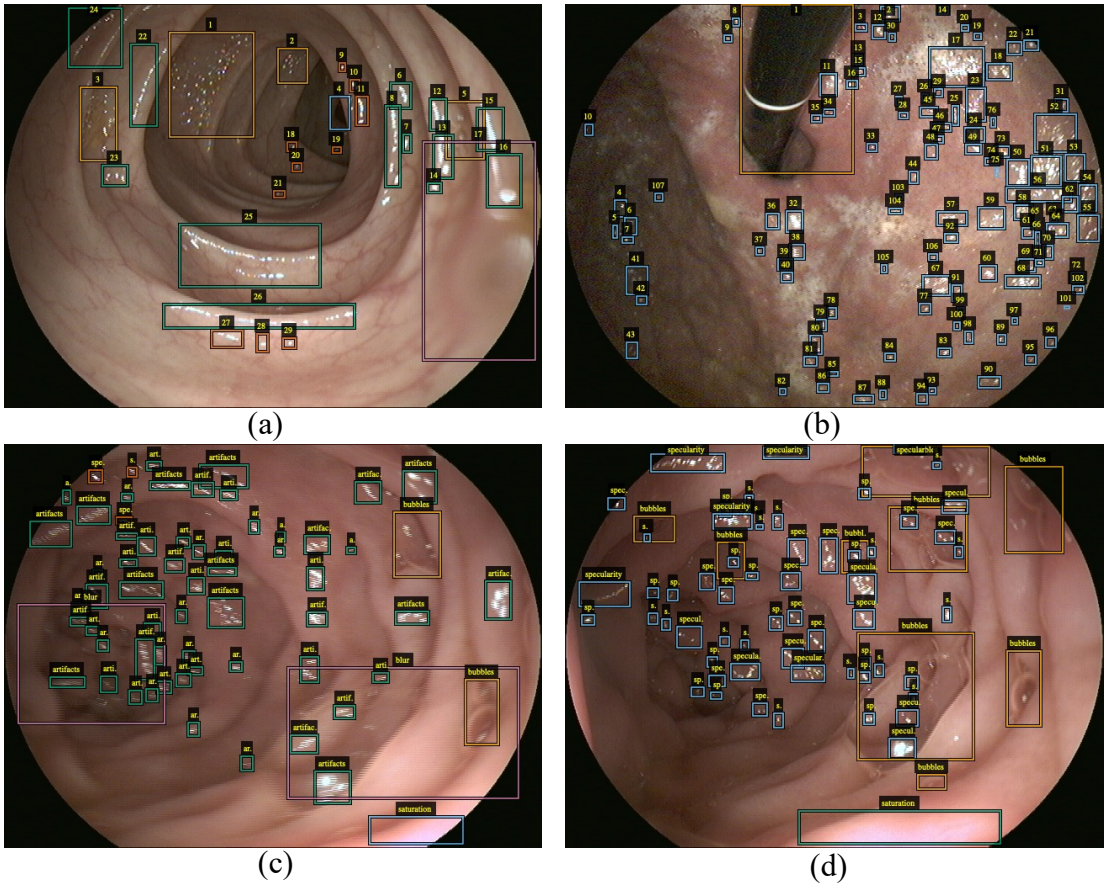


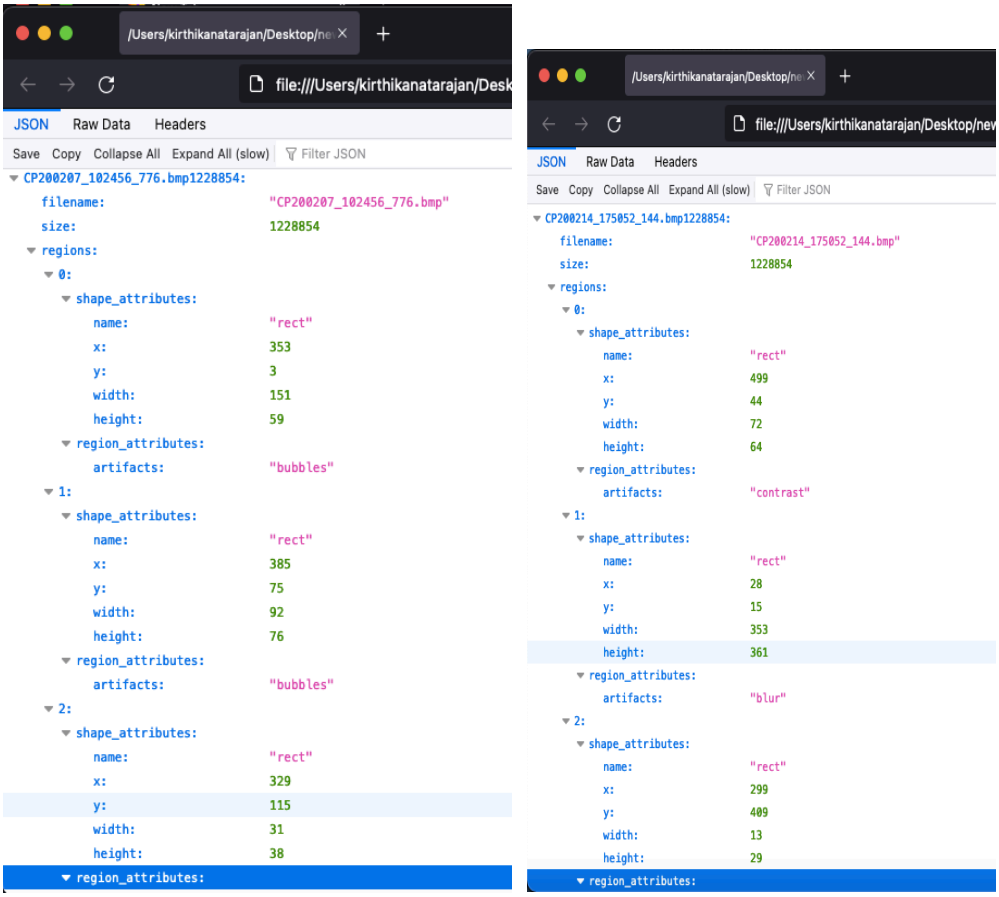
Figure 3.9 (a) - (d) Sample Annotated Images

A detailed walk-through of all annotated images will reveal that comparison on bounding boxes per artefact can never be equal due to the sparse distribution of few artefacts such as, contrast, bubbles, dense distribution of specular reflections and miscellaneous artefacts.

3.3.2 Annotation Software and Procedure

VGG Image Annotator (VIA), a free annotation tool provided by Oxford University, United Kingdom is used for all annotations (https://www.robots.ox.ac.uk). Figure 3.10 (a) & (b) show sample annotation files. Each artefact present in the endoscopic image is identified individually. Each artefact is localized using a precise bounding box. The dimensions of every bounding box varies based on the nature and size of artefact. Artefacts such as, instrument and saturation needs to be covered using a single bounding box. Artefacts such as, bubbles, specular reflections are to be addressed based on their nature in the frame that is considered. Hence the nature of bounding box for every artefact varies from frame to frame.

The coordinates of the bounding box plays a major role while utilizing the image for training deep learning based algorithms. The bounding box must not overestimate the artefact. Hence a bounding box must be drawn such that only the artefact affected area is covered. Enough care is taken to avoid overestimation. In general, all the bounding box used in this research follows either a square or a rectangle shape. The images from custom dataset will be combined with annotated images from EAD dataset. Hence to maintain uniformity all the artefacts are bounded using square or a rectangle. The tools gives a separate annotation file encapsulating the annotated data of every artefact with its corresponding coordinates mapped to the artefact. Hence images with corresponding annotation files are saved and created as separate annotated dataset.



(a) (b)

Figure 3.10 (a) and (b) Sample Annotation Files

Figure 3.11 demonstrates the process of crafting a new dataset. The process starts with collection of unlabelled endoscopic images affected by artefacts. The next step is to

label the images with help of clinician that results in final dataset containing endoscopic images labelled for artefacts.

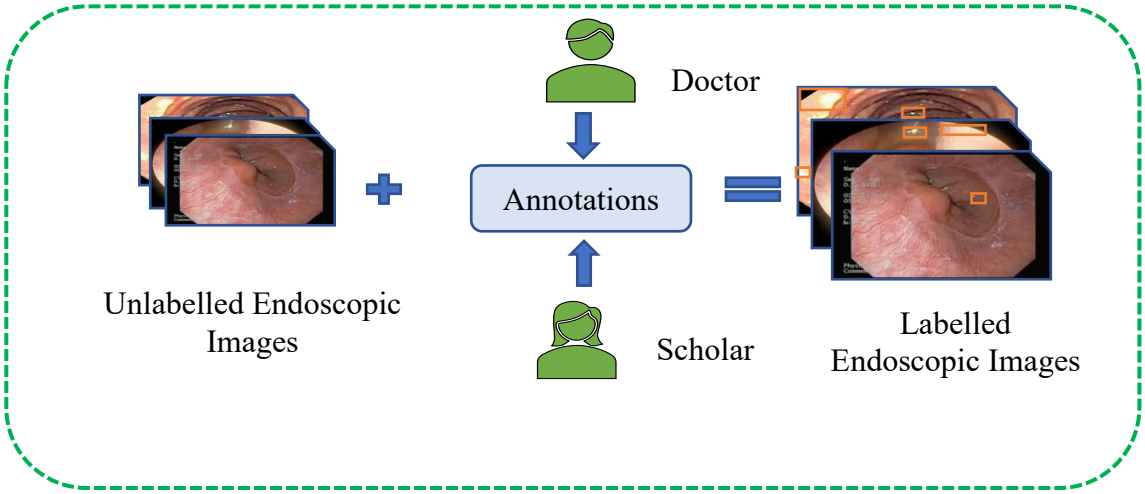


Figure 3.11 Process of Creating Custom Dataset

3.4 PARTITIONING OF DATASET

All three datasets are grouped and partitioned as shown in the Figure 3.12. First, the dataset is divided into images for training (80%) and testing (20%). Later, the training set is further split into the training set and the validation set. As a result, on average, 70% of images are used for training, 10% of the images are used for validation, and 20% of the images are used for testing.

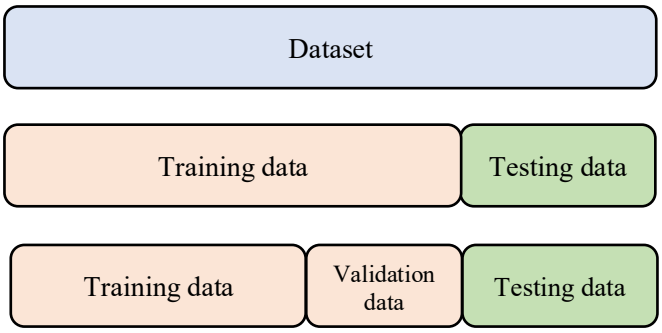


Figure 3.12 Partitioning of Dataset

Manually the datasets are fragmented into train, validation and test. The overall set contains ~4900, ~700 and ~1400 on each training, validation and test dataset. After partitioning the total number of images for training seemed trivial. Hence data augmentation technique is deployed to boost the dimensions of the dataset.

3.5 MANNUAL DATA AUGMENTATION FOR ARTEFACT DETECTION

The images for training constitute ~4900 images, as mentioned in the section 3.4. Hence the data augmentation is chosen to artificially expand the existing dataset to suit the DL model data requirement. DL algorithms crave for data. It is costly and time-consuming to collect millions of images and label them. Especially in medical imaging, the labelling must be done either by doctor or with the assistance of an expert. It is practically unviable for a health expert to dedicate so much time for labelling of images. Data augmentation techniques are developed to combat this issue. It extracts the most out of the dataset, where this assumption is not validated yet, through research findings. In data augmentation, images from the available dataset are chosen. Some alterations will be made to create a transformed version of the original images. The detector can be trained and be more robust when such techniques are deployed.

Standard data augmentation techniques include position augmentation and colour augmentation. The former includes rotation, padding, scaling, cropping, flipping, translation and affine transformation. The later consists of varying hue, saturation, brightness and contrast. Recently, many researchers proposed various augmentation techniques such as, colour jittering, edge enhancement, decolourizing and adding noise. Depending on the augmentation approach employed, the detector's performance vary.

For applications such as, medical imaging, data augmentation techniques can be chosen based on a literature study or by training the detector and evaluating its performance on a trial and error basis. In this research, manual data augmentation techniques such as, rotation and flipping are chosen based on the literature study. For rotating an image the following parameters are considered, the centre point of an image about which the image has to be rotated, and the width (w) and height (h) of the image. The size of output and the input image is set to be the same. The image can be rotated at various angles. The angle is initially converted to radians using the following formula given in Equation 3.1 where pi = 3.1415.

$$cal = pi * \frac{angle}{180} \tag{3.1}$$

After obtaining the angle in radians the middle row and the middle column of the image is identified by dividing the height and width by two. Later, it is necessary to calculate

and map each output image pixel with its corresponding input image pixel value. The values are calculated using the Equation 3.2. where x' and y' represents the output image pixel values.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{3.2}$$

When an image is rotated, the ground truth bounding box must necessarily be rotated. The values of ground truth bounding boxes are converted to standard cartesian coordinates for rotation. Angles for rotation like 90°, 180° and 270° are chosen. The formulas used to calculate the new coordinates are given from Equation 3.3 to 3.6 where x1, x2, y1 and y2 are the coordinate values and $x_{new}, y_{new,}\ w_{new}, h_{new}$ are the new coordinates.

$$x_{new} = \frac{(x1+x2)}{2*image_{width}} \tag{3.3}$$

$$y_{new} = \frac{(y1+y2)}{2*image_{height}} \tag{3.4}$$

$$w_{new} = \frac{x1-x2}{image_{width}} \tag{3.5}$$

$$h_{new} = \frac{y1-y2}{image_{height}} \tag{3.6}$$

All the rotated images are vertically flipped. Flipping, is an operation where the image is turned over a straight line as a result of which, it forms a mirror image. Flipping is accomplished using the Equation given in 3.7 and 3.8 where x1 and x2 are the coordinates.

$$x1 = -x1 \tag{3.7}$$

$$x2 = -x2 \tag{3.8}$$

As a result, after data augmentation, the dataset is inflated by eight times its original size. The discussed augmentation techniques are chosen based on trial and error basis. Figure 3.13 (a) embraces the actual image. Figure 3.13 (b) shows the actual image after 90° rotation. Figure 3.13 (c) shows image after 180° rotation and Figure 3.13 (d) show image rotated at 270° rotations.

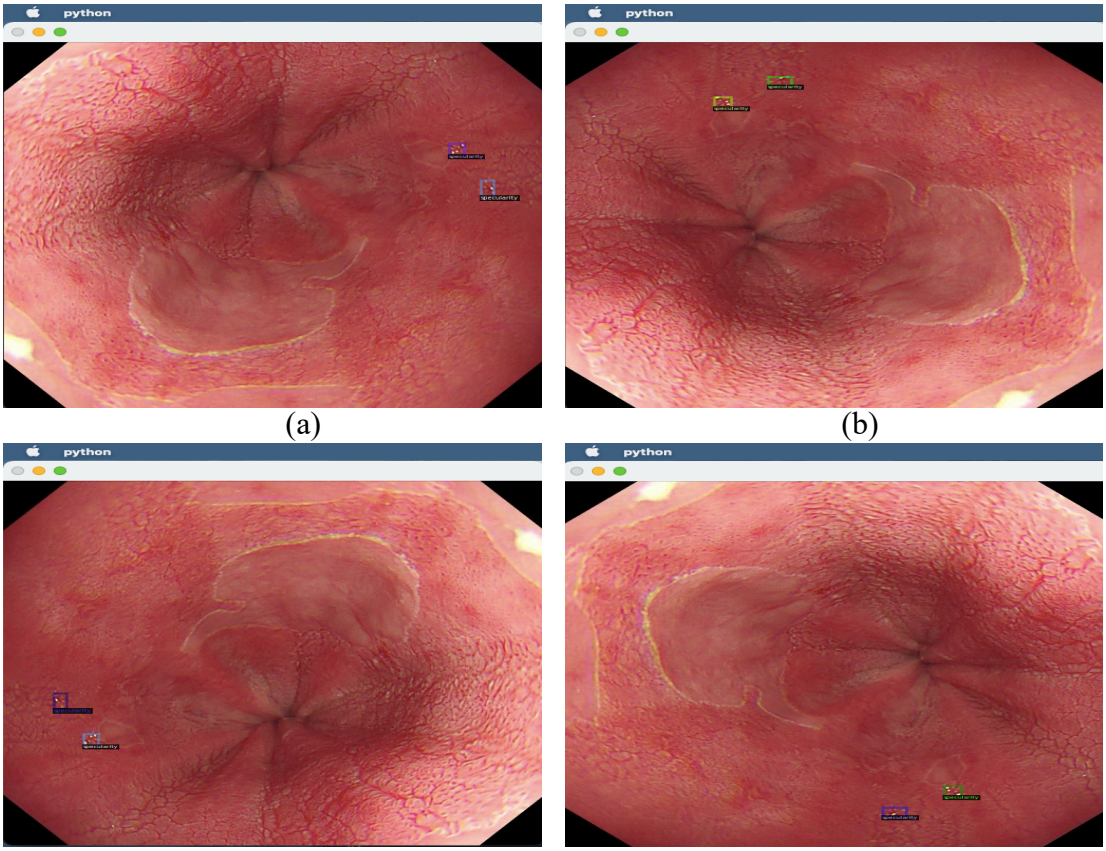


Figure 3.13 (a) - (d) Training Set Images Rotated at Various Angles

Figure 3.14 (a) illustrates the original image after flipping. Figure 3.14 (b) – (d) shows the flipped image that are already rotated at various degrees like 90°, 180° and 270°.

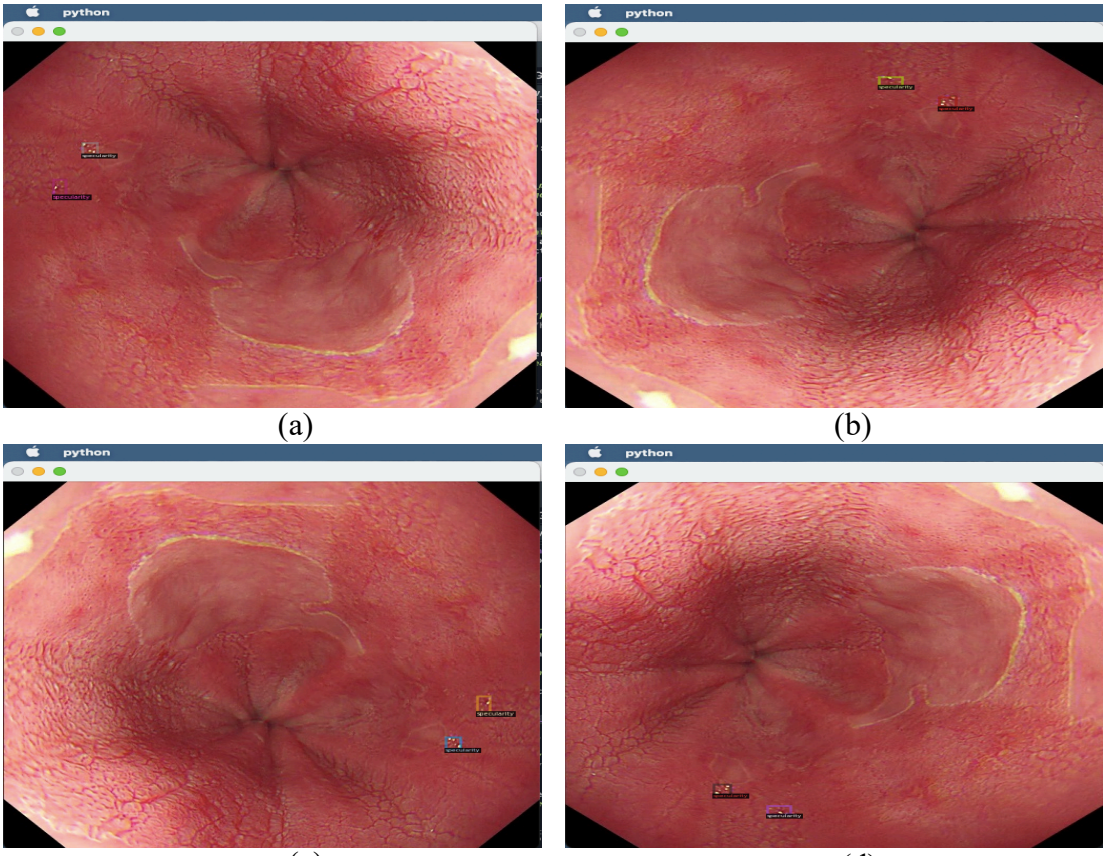


Figure 3.14 (a) - (d) Training Set Images Flipped after Rotating at Various Angles

3.6 RUN TIME DATA AUGMENTATION FOR ARTEFACT DETECTION

YOLOv3 is trained with the traditional data augmentation techniques such as, colour and position augmentation. The images that are to be augmented, and the method is chosen randomly.

Data augmentation techniques of YOLOv4 include colour augmentation and position augmentation. Along with this popular technique, two new techniques "cut-mix" and "mosaic" are introduced in YOLOv4. In cut-mix augmentation technique, two random images are combined through as follows:

Using the chosen random images two new images must be created after cut-mix. Let the new images be $\tilde{x}\ and\ \tilde{y}$. That is created by combining two training samples (x_A, y_A) and (x_B, y_B). The combining operation for cut-mix is given in Equation 3.9 and 3.10.

$$\tilde{x} = M \odot x_A + (1 - M) \odot x_B \quad (3.9)$$

$$\tilde{y} = \lambda y_A + (1 - \lambda) y_B \quad (3.10)$$

Where M $\epsilon \{0, 1\}^{W*H}$ denotes a binary mask. It indicates the portion of image that has to be filled from other sample input image. W and H represent the width and height of the original image. 1 represents a binary mask filled with ones. In Equation 3.9 $\odot$ operator represents elementwise multiplication operation. λ is the combination ratio. It is sampled from a uniform distribution (0,1). Let B=(r_x, r_y, r_w, r_h). B indicates the cropping region. B region in the image x_A is removed and filled up in x_B. r_x, r_y, r_w, r_h are calculated using the following equation.

$r_x \sim Unif(0, W)$, $r_y \sim Unif(0, H)$, $r_w = W\sqrt{1-\lambda}$, $r_h = H\sqrt{1-\lambda}$. The cropping ratio is given by $\frac{r_w . r_h}{W.H} = 1 - \lambda$. Figure 3.15 explains cut-mix augmentation on an endoscopic image.

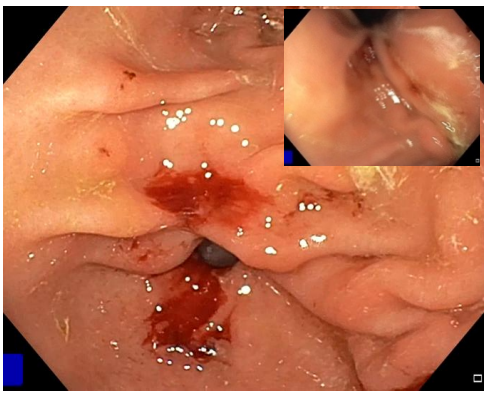


Figure 3.15 Cut-mix Data Augmentation in YOLOv4

YOLOv4 comes with another data augmentation technique called mosaic. It is an improvement over cut-mix data augmentation algorithm. Mosaic data augmentation technique pools four images from the training dataset. It combines four images into single image. The augmentation algorithm resizes the input images into equal sized grid. All four images are stitched after resizing. A random cut out of the stitched image becomes the mosaiced image. The final images consists of a portion from each of the four images considered. This algorithm allows the model to learn new features. This algorithm has the capability to identify objects that are at smaller scale. Figure 3.16 shows a mosaic augmented endoscopic image.

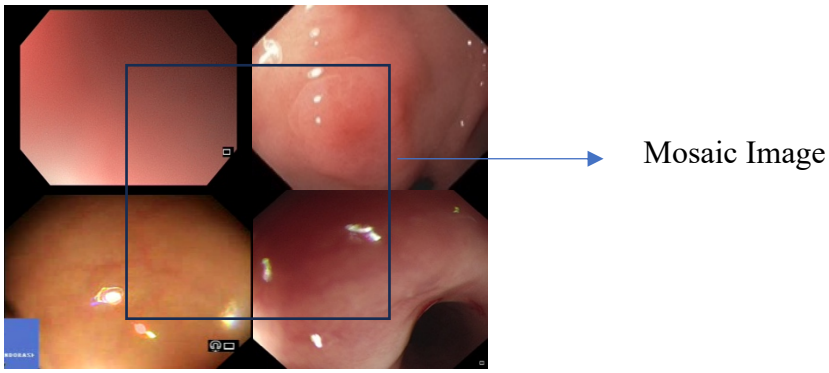


Figure 3.16 Mosaic Data Augmentation in YOLOv4

3.7 SYSTEM CONFIGURATION

The manual data augmentation process is coded in python programming language using OpenCV and numerical python library. The system configuration includes a 1.3GHz dual-core intel i5 processor with 4 Giga Bytes (GB) Double Data Rate version 3 (DDR3) Random Access Memory (RAM) and Intel High Definition (HD) Graphics 5000 series card.

3.8 STATEMENT OF ETHICS

Collecting the endoscopic images from the hospitals requires Institutional Human Ethics Committee (IHEC) clearance. The process has been done, and clearance certificate has been issued by the IHEC team to carry out research activities.

3.9 CHAPTER SUMMARY

The benefits and drawbacks of the EAD public dataset are outlined in this chapter. EAD2020 is just an extension of EAD2019. Many missing annotations, repetition of frames are rectified in EAD2020. Due to vast need of images, a brand new dataset is created. The dataset holds 2400 images. The public and custom dataset images are pooled and distributed for training and testing. The train set for detection task is manually data augmented using augmentation techniques like rotation at various angles such as, 90°, 120° and 270° and vertical flipping. The custom dataset created serves for artefact detection. Employing images from both datasets, all models are trained and tested. Few other run time augmentation techniques such as, colour manipulation, cut-mix and mosaic are employed during training for artefact detection. The combined dataset has more images with artefacts such as, instrument, blur and contrast. Hence the model has more additional to learn.