

DNN MODELS AND POSTPROCESSING THRESHOLDS FOR ENDOSCOPY ARTIFACT DETECTION IN PRACTICE

Seiryu Watanabe^{1,2}, Shigeto Seno¹, Hideo Matsuda¹

¹Department of Bioinformatic Engineering, Osaka University, Japan

²Autonomous Mobile Systems Laboratory, Meiji University 1-1-1 Higashimita, Japan

ABSTRACT

We tackled the problem of multi-class artifact detection and segmentation in endoscopic video frames using different deep learning algorithms and strategies. In particular we proposed to combine the advantages of two state-of-the-art deep object detection algorithms, YOLOv3 and Mask R-CNN. With Mask R-CNN we achieved improved performance by leveraging the available image segmentation annotations to aid bounding box detection. Appropriate thresholding of objectness score and non-maximum suppression (NMS) were subsequently applied to achieve high leaderboard test scores.

Index Terms— Endoscopic artefact detection, Mask R-CNN, YOLOv3

1. INTRODUCTION

This paper details the work done taking part in the Endoscopic artefact detection challenge (EAD2019) [1, 2] on all three sub-challenge, artefact detection (task #1), segmentation (task #2) and generalization (task #3). Specifically we investigated i) the use of Mask R-CNN to use segmentation to improve the quality of artefact bounding box detection, ii) the combination of single-stage YOLOv3 and two-stage Mask R-CNN bounding box predictions and iii) the tuning of post-processing thresholds to reduce false positive detection in practical applications.

2. METHODS

2.1. Segmentation Aided Artefact Detection

The provided segmentation training data was composed of around 589 images of which 498 were semantically annotated for 5 artefact classes, specularity, image artifact, bubbles, saturation and instrument as binary masks. After removal of duplicate masks, we get 474 uniquely annotated segmentation video frames. Bounding box annotation for artefact detection consisted was released in two phases, 886 images in training data I and 1306 images in training data II. We first trained a deep convolutional neural network (DNN) using the 474 segmentation images. A total of 3312 individual cropped im-

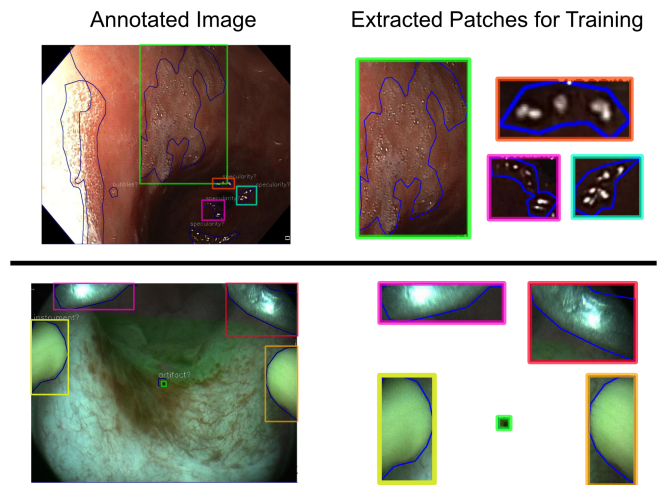


Fig. 1: Example images (left) and extracted image patches after cropping (right) from the provided segmentation dataset of EAD2019.

aged patches from the available segmentation masks cropped by the width and height of each mask region, Fig. 1 were then extracted. The extracted patches were then used to refine the previously trained DNN. This gave us a network that could quickly assess if candidate regions contained artifacts in the training data. We then created an augmented datasets by adding the 474 images of the segmentation dataset to the provided 2186 images with bounding box annotations in the detection challenge. To further increase the number of training images overall we carried out data augmentation, rotating each image three times at increments of 90^0 angles as well as horizontal flipping. This gave a total of $(474 + 2186) \cdot 8 = 21280$ training images. For image patches with bounding box annotations without segmentation masks, we generated corresponding segmentation masks using the patch trained DNN, Fig. 2. We then trained a Mask R-CNN [3] model with a Feature Pyramid Network and ResNet101 backbone on the constructed augmented image dataset.

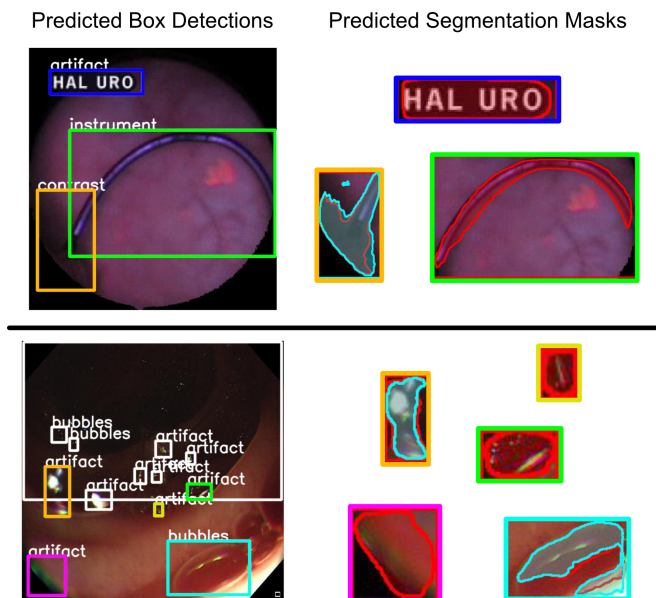


Fig. 2: Example of images with only bounding box annotations provided (left) and corresponding generated segmentation masks for each extracted patch (right) using a DNN trained on images with provided segmentation masks.

2.2. Combining Mask R-CNN and YOLOv3 for artefact detection

While Mask R-CNN uses a Region Proposal Network (RPN) to get high accuracies, this loses the spatial relation of object and non-object regions. Specularity, Saturation, Artifact, Bubbles, and Instruments are kinds of objects and have clear defined image boundaries but Blur and Contrast artefacts are image areas that do not have clear boundary. We thus used YOLOv3 [4] for blur and contrast artefact detection, Fig. 3. Unlike Mask R-CNN, YOLOv3 splits the image into several spatial grids and predicts bounding boxes and class probabilities based on the grid thus retaining the spatial relation of objects and background. The test dataset was thus processed with both Mask R-CNN and YOLOv3 models with the result of Mask R-CNN predicted blur and contrast replaced with the corresponding result of YOLOv3. By doing this, IoU was increased 13% and mAP increased 2% compared to the result without integrating YOLOv3.

2.3. Postprocessing thresholds for detection and segmentation

Now we have several bounding boxes, masks, class probabilities and mean average precision (mAP) for each classes. Normally the research process ends at this point when we care only about a good model for artifact detection. However for this competition the rules of includes Intersection over Union (IoU) as an additional score for object detection, 0.6 mAP + 0.4 IoU which requires tuning of detection thresh-

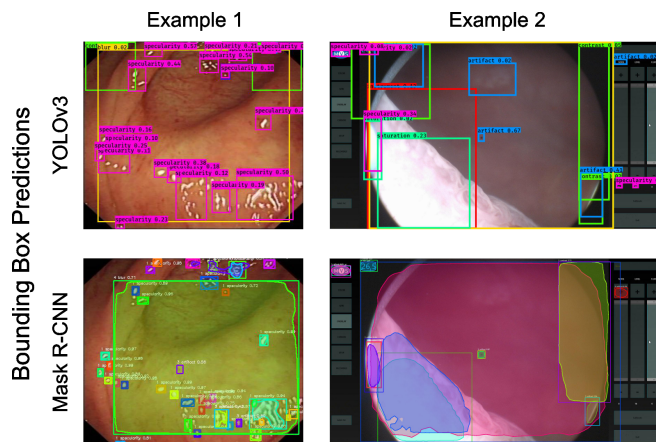


Fig. 3: Example bounding box detections using YOLOv3 (top) and Mask R-CNN (bottom). For Mask R-CNN predicted segmentation masks for each bounding box is additionally visualised.

olds. This means one not only has to reduce the number of false positives (FP) and increase the number of true positive (TP) where a positive match covers at least one quarter of the ground truth area for mAP but additionally must care about how precisely positive areas are detected and how much of the true are is covered. Similarly the segmentation score uses Jaccard, Dice and F2-score for the binary mask of each class so a strict threshold is needed to reduce FP areas regardless of their probabilities. These metrics are good for practical application. During an endoscopy examination the ideal algorithm should not remove crucial areas such that reducing FP is more important than reducing FN. We thus investigated different thresholds for detection and segmentation. In the end we a NMS (non-max suppression) threshold of 0.5 for detection because the object should not overlap with other objects of the same score and an objectness threshold of 0.01 to balance the mAP and IoU scores. IoU was increased $\sim 10\%$ while mAP was decreased $\sim 1\%$ by applying NMS. We used 0.5 as a score threshold for segmentation task and did not apply non maximum suppression (NMS) because the same image region can be labelled for several classes.

3. CONCLUSIONS

We show that the provided segmentation train dataset was good enough to produce segmentation masks for images with only bounding box annotations. The quality of the constructed dataset was sufficient to improve DNN models across every artifact detection metric. We show that Mask R-CNN has good ability to locate endoscopic artefacts with good image boundaries while YOLOv3 was better for locating artefacts with unclear boundary area such as blur and contrast. We found careful setting of thresholds on objectness score and NMS steeply increased IOU with small deficit in

mAP.

4. REFERENCES

- [1] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnires, Victor Loschenov, Enrico Grisan, Walter Blondel, and Jens Rittscher, “Endoscopy artifact detection (ead 2019) challenge dataset,” 2019.
- [2] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James East, Xin Lu, and Jens Rittscher, “A deep learning framework for quality assessment and restoration in video endoscopy,” 2019.
- [3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, “Mask r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
- [4] Joseph Redmon and Ali Farhadi, “Yolov3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.