ENDOSCOPIC ARTEFACT DETECTION AND SEGMENTATION WITH DEEP CONVOLUTIONAL NEURAL NETWORK

Suhui Yang, Guanju Cheng

Ping An Technology (Shenzhen) Co. Ltd., Shenzhen, China

ABSTRACT

Endoscopic artefact detection challenge (EAD2019[**?**]) includes three tasks: (1) Multi-class artefact detection: localization of bounding boxes and class labels for 7 artefact classes for given frames (specularity, saturation, artefact, blur, contrast, bubbles and instrument); (2) Region segmentation: precise boundary delineation of detected artefacts (instrument, specularity, artefact, bubbles and saturation); (3) Detection generalization: detection performance independent of specific data type and source. We participated all three tasks of EAD2019, and this manuscript summarizes our solution based on deep learning for each task. In short, for task 1, we apply the improved Cascade R-CNN [1] model combined with feature pyramid networks (FPN) [2] to deal with multi-class artefact detection; for task 2, we apply the network architecture like Deeplab v3+ [3] with different backbones (ResNet101 [4] and MobileNet [5]) to segment multi-class artefact regions; for task 3, we used Cycle-GAN [6] and then perform image translation between training dataset and testing dataset to improve the model generalization of multi-class artefact detection. Besides, we apply unsupervised t-SNE [7] to visualize the date distribution to achieve targeted data reduction and augmentation before training detection and segmentation model; and finally, some effective strategies of model fusion and post-processing are also used to obtain the final results.

Index Terms— Endoscopic artefact detection challenge, t-SNE, cascade R-CNN, generalization

1. INTRODUCTION

Endoscopy is a widely used clinical procedure for the early detection of numerous cancers (e.g., nasopharyngeal, oesophageal adenocarcinoma, gastric, colorectal cancers, bladder cancer etc.), therapeutic procedures and minimally invasive surgery (e.g., laparoscopy). EAD2019 challenge[**?**] proposal aims to address the following key problems inherent in all video endoscopy: 1) Multi-class artefact detection: Existing endoscopy workflows detect only one artefact class which is insufficient to obtain high-quality frame restoration. In general, the same video frame can be corrupted with multiple artefacts, e.g. motion blur, specular reflections, and low contrast can be present in the same frame. Further, not all artefact types contaminate the frame equally. So, unless multiple artefacts present in the frame are known with their precise spatial location, clinically relevant frame restoration quality cannot be guaranteed. Another advantage of such detection is that frame quality assessments can be guided to minimise the number of frames that gets discarded during automated video analysis. 2) Multi-class artefact region segmentation: Frame artefacts typically have irregular shapes that are non-rectangular and consequently are overestimated by the detected bounding boxes. Development of accurate semantic segmentation methods to precisely delineate the boundaries of each detected frame artefact will enable optimized restoration of video frames without sacrificing information. 3) Multi-class artefact generalisation: It is important for algorithms to avoid biases induced by specific training data sets. Also, it is well known that expert annotation generation is time consuming and can be infeasible for many data institutions. In this challenge, we encourage the participants to develop machine learning algorithms that can be used across different endoscopic datasets worldwide based on our large combined dataset from 6 different institutions.

2. MATERIALS AND METHODS

2.1. Task 1: Multi-class artefact detection

EAD2019 [8, 9] provides two batches of training data for multi-class artefact detection, the first batch contains 886 endoscopic images labeled with 9352 bounding boxes and the second batch contains labeled 1306 endoscopic images labeled with 8466 bounding boxes. After checking the training data, we notice that there may be two difficulties in this task. One is unbalance sample distribution, another is various size/aspect ratio of image and detection object. As shown in table 1, there are 4074 specularity and only 327 blur in training data1, and there are 3487 artefact and only 46 instrument in training data2.

Based on this observation, we therefore propose an improved Cascade R-CNN [1] as our detection model (Figure 1). Compared to original Cascade R-CNN, we add the FPN [2] module during feature extraction. As shown in Figure 1, there are two main sub-modules, including multi-scale feature

num Classes	First batch of training data(886)	Second batch of training data(1306)
specularity	4074(44%)	1761(21%)
saturation	511(44%)	611(7%)
artefact	1609(17%)	3487(41%)
blur	327(%)	348(4%)
contrast	686(7%)	872(%)
bubbles	1738(19%)	1341(16%)
instrument	407(4%)	46(%)
total	9352	8466

Table 1. Statistics of two batches of training data

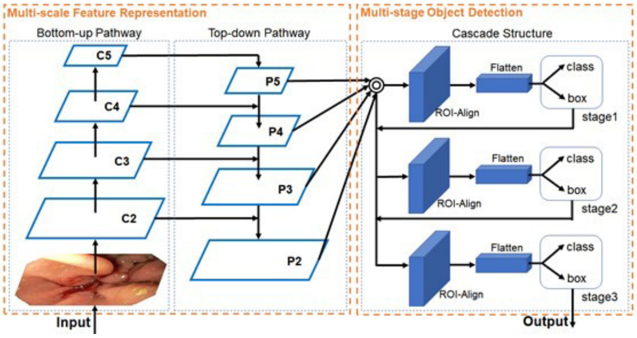


Fig. 1. The flowchart of our detection network, a improved Cascade R-CNN by adding FPN module.

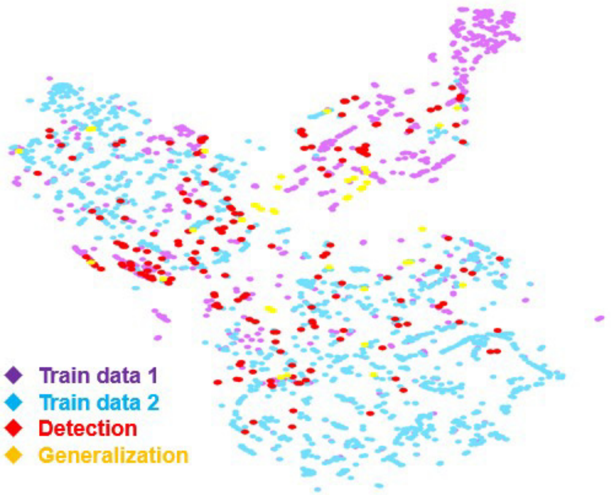


Fig. 2. Data visualization with t-SNE for training data1, training data2, validation data for detection (task 1) and generalization (task 3).

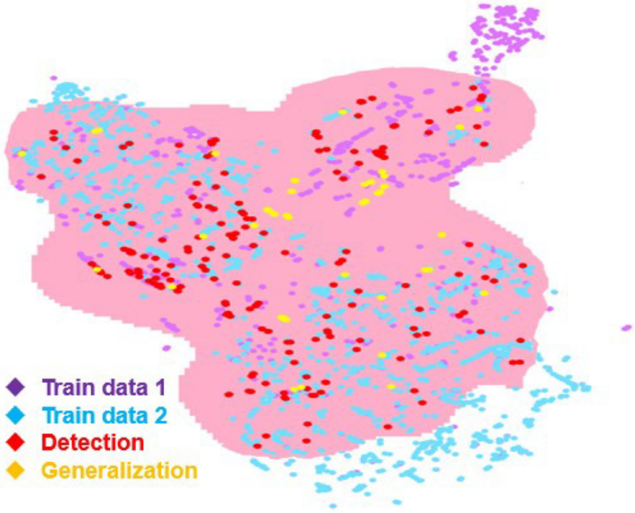


Fig. 3. According to the distribution of testing data for detection task and generalization task, the training data within the shaded area were selected to feed the model, which ignored the noisy outliers have less contribution for model training.

representation and multi-stage object detection with cascade structures. The module of multi-scale feature representation consists of a bottom-up pathway and a top-down pathway [2]. Using ResNet101 [4] as backbone, the input image is processed through bottom-up pathway with a series of residual blocks. We denote the feature activation outputs of last residual blocks as $\{C2, C3, C4, C5\}$, and we do not include the output of first residual block due to memory space. Then in top-down pathway, each feature map is constructed by merging the corresponding bottom-up map and the unsampled map from a coarser-resolution feature map with a factor of 2. The final feature maps are denoted as $\{\mathrm{P2, P3, P4, P5}\}$ with different spatial sizes corresponding to $\{\mathrm{C2, C3, C4, C5}\}$. With the FPN, we could produce the multi-scale feature representations, which can improve the detection rate of small objects (e.g. specularity) by combining low-level features and high-level semantic information. In general, IoU and mAP is a pair of mutually contradictory index, e.g. an object detector with higher IoU value may usually produce noisy detections leading to low mAP. We apply three cascade stages of object detection networks (R-CNN) to improve the performance [2]. This structure can prevent mAP from dropping sharply when IOU is high between the prediction box and the real box. Besides the network architecture, we pay more attention in data distribution. We apply an unsupervised nonlinear dimensionality reduction method called t-SNE [7] to visualizing all the dataset including training data1, training data2, validation data for detection and generalization (shown in Figure 2). We find there are different data distributions among two training datasets and validation data. Therefore, we delete some outliers and continuous frames in training data2, and then perform data augmentation for categories with fewer sample (e.g. saturation and blur). Similar operation is also carried out for task 2 and task 3.

Methods	mAP$_d$	IoU$_d$	score$_d$
Faster RCNN[10]	0.2618	0.3448	0.2950
Cascade RCNN[1]	0.2996	0.3221	0.3086

Table 2. Results of different models

2.2. Task 2: Region segmentation

We select Deeplab v3+ [3] network for multi-class artefact segmentation, with different backbones (ResNet101 [4] and MobileNet [5]). After backbone network, we add 5 parallel convolution layers as the feature extraction layers, which include one 1*1 convolutional layer, three 3*3 dilated convolutional layers with different ratios of 6, 12, 18, and one global pooling layer. Then these feature maps are merged and unsampled to achieve the region segmentation.

2.3. Task 3: Detection generalization

For detection generalization, we translate the training data to the style of validation data with Cycle-GAN [6]. we replace the deconvolution with the linear interpolation with 1*1 convolution to improve the performance of style transfer. Then we retrain the detection model with translated training data and test its performance..

3. EXPERIMENTS AND RESULTS

In our experiments we evaluated the method of each task in detail. We also compare the experimental results of our methods and Faster-rcnn [10] model for task 1.

3.1. Task 1: Multi-class artefact detection

We use SGD method to optimize the improved Cascade R-CNN. The learning rate is 0.005 with a staged decline mode, and a total of 30 epoch is performed, and the batch size is 2, all images are resized to 1333*800. Table 1 shows the results with different methods. Table 2 shows the results with different data methods. In validation data, we perform the data augmentation with the operations of flip and contrast. Note that the best results for each condition are achieved by the technique of non-maximum suppression (NMS[11]). As shown in Table 2, the cascade-rcnn method achieves a good trade-off between mAP and IOU, which is 1.36% higher than the faster-rcnn [10] model. We also compared the results of the different methods in detail, As shown in the table 3, the final result of model7 brought 4.83% improvement when compared to the original cascade-rcnn results.

3.2. Task 2: Region segmentation

The Adam optimizer is used, the initial learning rate is 0.007, a total of 30k iterations are trained, and the batch size is 10., all the images are resized to 513*513.We select multi-class sigmoid as loss function since there may be overlap among different artefact classes. Table 4 shows the results of multi-class artefact segmentation by different methods. Due to limited training datasets, the contours of some instruments can't be extracted completely, to remedy the over-segmentation problem of instruments, we combined a marker-based watershed segmentation, which took partially extracted regions as markers, to perceptually group together the mulitiple parts of instruments. In this way, the multiple parts of instruments are perceptually merged according to their regional homogeneity. And the result of segmentation is shown in Figure 4 which we can see how the post-processing improves the segmentation results clearly. Besides, from Table 4, the result of merging the two backbones (Resnet101[4], Mobilenet[5]) Increased from 0.6414 to 0.6568, with an increase of nearly 1.5 percentage points, and further improved to 0.6700 by post-processing such as regional growth.

Methods	mAP$_d$	IoU$_d$	score$_d$
Model1(only training data1)	0.2210	0.4504	0.3127
Model2(only training data2)	0.2138	0.4323	0.3012
Model3(training data1+ training data2)	0.2996	0.3221	0.3086
Model4(selected by t-SNE)	0.2379	0.4512	0.3235
Model5(selected by t-SNE + data augmentation for training data)	0.2658	0.4476	0.3385
Model6(selected by t-SNE + data augmentation for testing data)	0.2633	0.4663	0.3445
Model7(model5 and model 6 fusion by NMS)	0.3235	0.4172	0.3610

Table 3. Results of different datasets by the improved Cascade R-CNN

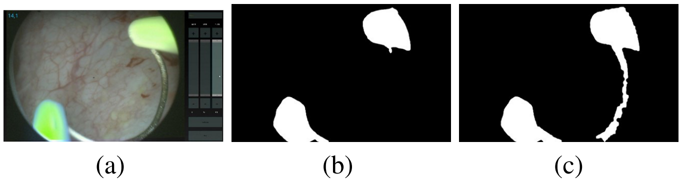

(a) (b) (c)

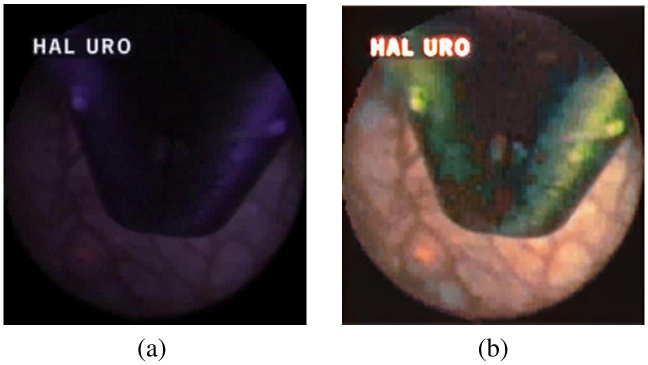
Fig. 4. (a) the original image; (b) the results of model without post-processing; (c) the final result.

3.3. Task 3: Detection generalization

We trained the Cycle-GAN with some hyper parameters: the Adam optimizer is used, the initial learning rate is 0.002, a total of 30 epochs are trained, and the batch size is 1, all

Methods	Overlap	F2-score	score$_s$
Resnet101 backbone	0.6288	0.6795	0.6414
Ensemble two backbones	0.6592	0.6937	0.6568
Ensemble +Post-processing	0.6612	0.6964	0.6700

Table 4. Segmentation results of by different methods



(a)

(b)

Fig. 5. (a) the original image; (b) the translated image.

Methods	mAP$_g$	dev$_g$
Train with original training data	0.3187	0.1018
Train with translated training data by Cycle-GAN[6]	0.3747	0.0693

Table 5. Results of detection generalization

the images are resized to 512*512. Then trained the detection model in task 1 with original and translated training data respectively, and compare their performance of detection generalization. As shown in Table 5, the model trained with original training data obtains mAP_g=0.3187, and dev_g=0.1018, while the model trained with translated data obtains mAP_g=0.3747, and dev_g=0.0693. Therefore, the performance of detection generalization improved with style transfer.

4. CONCLUSION

In Task 1, the better results are obtained by combining the fpn and cascade-rcnn models. The mAP and IOU evaluation indicators are more balanced. At the same time, using t-SNE to automatically select similar samples between the two batches of training data and the test datasets, which is helpful to accelerate the model training; In the task 2, the deeplabv3+ model is supplemented by the multi-class sigmoid loss function to improve the segmentation effect of the model; In the task 3, We presented using cycle-gan to translate the training set of task 1 to the testing set in task 3, and the fine-tuning the detection model of task 1, which could effectively improve the generalization of the detection model of task 1.

ACKNOWLEGMENT

We gratefully thank Dr. Peng Gao, Dr. Bin Lv, Dr. Jun Wang, Xia Zhou, Ge Li, Chenfeng Zhang and Yue Wang for their valuable discussion and kind support during this challenge.

5. REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos, “Cascade r-cnn: Delving into high quality object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

[2] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, “Feature pyramid networks for object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 801–818.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[5] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[7] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[8] 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,” *CoRR*, vol. abs/1905.03209, 2019.

[9] 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," *CoRR*, vol. abs/1904.07073, 2019.

[10] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[11] Alexander Neubeck and Luc Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006, vol. 3, pp. 850–855.