Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.jzus.zju.edu.cn; www.springerlink.com
E-mail: jzus@zju.edu.cn

Correspondence:

Detection and segmentation of multi-class artifacts in endoscopy

Yan-yi ZHANG[1], Di XIE[†‡2]

[1]*Department of Psychology, the Children's Hospital, School of Medicine, Zhejiang University, Hangzhou 310003, China*

[2]*Hikvision Research Institute, Hangzhou 310052, China*

[†]E-mail: xiedi@hikvision.com

https://doi.org/10.1631/jzus.B1900340

Endoscopy may be used for early screening of various cancers, such as nasopharyngeal cancer, esophageal adenocarcinoma, gastric cancer, colorectal cancer, and bladder cancer, and performing minimal invasive surgical procedures, such as laparoscopy surgery. During this procedure, an endoscope is used; it is a long, thin, rigid, or flexible tube having a light source and a camera at the tip, which facilitates visualization inside the affected organs on a screen and helps doctors in diagnosis.

Artifact detection in endoscopic imaging is not only essential but also poses a significant challenge to accurately diagnose hollow organ diseases. Artifacts can be classified into many categories. In endoscopic environments, there are several variants of artifacts that may be a cause of concern, including pixel saturations, motion blur, defocus, specular reflection, bubbles, and artificial devices (Fig. 1). The importance of precisely detecting these artifacts is essential to perform high-quality endoscopic frame restoration and crucial for realizing reliable computer-assisted endoscopy tools for improved patient care. Since artifacts are superimposed on the organ surfaces and have a great impact on the diagnosis, they may also interfere with other post-processing steps (video stitching and video frame retrieval).

Because most artifacts have irregular edges, they require two kinds of tasks depending on the different granularities: artifact detection and artifact region segmentation. In this paper, for the detection task, we extended Faster region-based convolutional neural network (R-CNN) (Ren et al., 2015) into a multi-stage cascaded R-CNN framework (Cai and Vasconcelos, 2018). The purpose of cascaded R-CNN is to train a series of bounding-box regressors which could improve the intersection over union (IoU) between the candidate outputs and the ground-truth boxes. Furthermore, cascaded R-CNN also considers the high negative correlation between detector performance and IoU under a given threshold (higher threshold raises the requirement of IoU and results in lower detection performance). In this paper, we have used a phased approach to gradually raise the IoU threshold when training cascaded R-CNN.

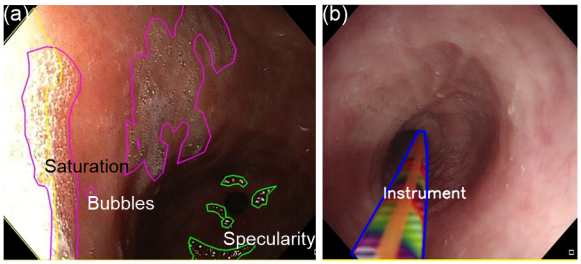

Fig. 1 Artifact categories in endoscopy artifact detection (EAD) dataset (Ali et al., 2019)

The yellow border regions are with high saturation, the purple border regions represent the bubbles, the green border regions represent the mirror reflection (a), and the blue border region represents an artificial device (b)

To resolve the multi-label issue in the segmentation task, we constructed a classifier chain (CC) method (Read et al., 2011), in which the input of each classifier is the original image along with the outputs of all previous classifiers. The "one-vs-rest" training strategy is incapable of modeling the internal association between labels; on the contrary, our proposed CC strategy could elucidate the relationship between different label types to a certain extent and address this issue.

Several studies have reported on the various detection and segmentation performance of medical

‡ Corresponding author
ORCID: Di XIE, https://orcid.org/0000-0001-8065-5901
© Zhejiang University and Springer-Verlag GmbH Germany, part of Springer Nature 2019

devices using endoscopy (Bouget et al., 2017; Münzer et al., 2018). Early detection and segmentation of medical devices typically depend on the unique color or shape of devices, particularly resorting to circle detection (Münzer et al., 2013). However, the accuracy attained is not reliable when the medical devices deal with complex conditions such as blurs, smoke fog, blood immersion, and bubbles. Classic support vector machine (SVM) and deformable part model (DPM) have also been introduced into endoscopic scenes to improve detection (Bouget et al., 2015). With the rise of deep learning in recent years, data-driven feature representation learning has demonstrated better performance on numerous tasks than manual design features. However, due to the lack of labeled endoscopic datasets, deep learning methods could not be applied regularly.

By considering the increasing amount of endoscopic data, a number of deep models performing well in other visual tasks, gradually become the mainstream processing methods for such images. In the 2017–2018 segmentation competition for medical devices (Allan et al., 2019), deep models such as UNet and fully convolution neural network (FCN) overtook the most competitive methods (Shvets et al., 2018). Unlike previous endoscopic scene datasets, endoscopy artifact detection (EAD 2019) detects and segments not only medical devices, but also areas of over-exposure, reflection, and special bubbles (Ali et al., 2019). Furthermore, in the segmentation task, one single pixel could be labeled for multiple categories, which is rather different from segmentation tasks previously. Our proposed approach achieves better results than previous methods in dealing with these new challenges.

Except for general artifact analysis, specific methods have been proposed by several researchers. Kalalembang et al. (2009) used various size block-based discrete cosine transform (DCT) calculations on the distorted image to detect blurring. Loukas and Georgiou (2015) performed smoke detection to ensure successful video retrieval. Leibetseder et al. (2017) used pre-trained CNNs and threshold-based saturation analysis for the smoke detection task.

In the multi-class artifact detection task, it is possible that we may encounter the problem of domain gaps between datasets; alternatively, the data have been derived from videos recorded using different endoscopes (Fig. 2).

Fig. 2 demonstrates different categories and sizes of ground-truth bounding-boxes, which indicates that there is a profound discrepancy among the scales of different classes. Additionally, objects from different classes may appear similar, e.g., red boxes and blue boxes in Fig. 2b; thus, it overlaps in some areas, which also means that even the manual labeling standards of one single object class may not be the same in different pictures.

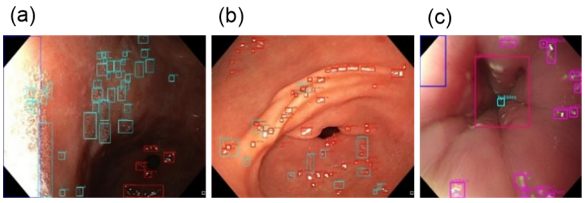

Fig. 2 Sample images of different domains
(a) and (b) are obtained from the same endoscopic device, while (c) is obtained from a different endoscope

To resolve the abovementioned problems, we proposed a deep neural network structure combining feature pyramid network (FPN) and cascaded R-CNN (Fig. 3). To detect small object areas, we used input images with 1024×1024 resolution; various data augmentation techniques were adopted to avoid overfitting due to the small amount of training data. Meanwhile, we introduced the FPN to extract and fuse multi-scale feature map information in a "top-down" mode, which generated a large number of high-quality candidate boxes through the regional proposal network (RPN), and significantly raised the detection recall rate. Finally, the cascade structure designed by us further improved the precision of candidate detection boxes via a layer-by-layer correction mechanism.

In comparison with single-scale detection methods (including Faster R-CNN), the proposed FPN structure could generate a large number of candidate boxes (Fig. 4), which greatly increased the recall rate in mean average precision (mAP). Although increased numbers of incorrect boxes would simultaneously reduce the precision, which may result in a lower mean IoU (mIoU); however, the overall performance could be improved by adjusting the output confidence and non-maximum suppression (NMS) threshold.

In data augmentation, we introduced noise interference to hue-saturation-value (HSV) color space and randomly alternated the channels of red-green-blue (RGB) image, which enlarged the domain scope

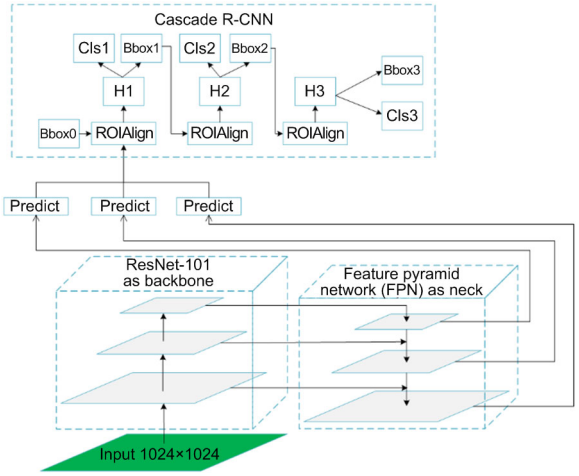

Fig. 3 Architecture of artifact detection network
All the regional proposal phases, including Bbox1, Bbox2, and Bbox3, are refined with increasing intersection over union (IoU) value, for the motivation of retaining proposals with low IoU but high quality. R-CNN: region-based convolutional neural network

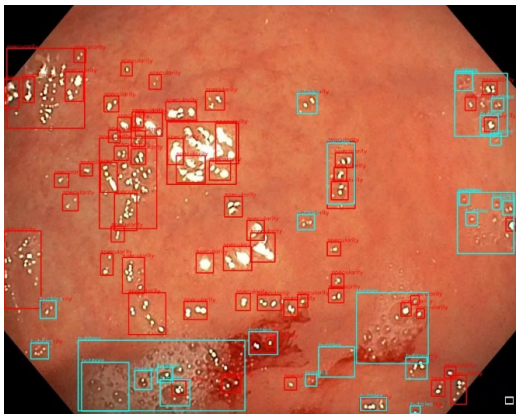

Fig. 4 Feature pyramid network (FPN) detection results
There are boxes within boxes, which show false positives

by introducing color divergence. Randomly cropping or enlarging an image, which simulates local focusing, was used as well to counteract the effects of a difference in scale. We also performed the geometric transformation on the images, including horizontal and vertical flips, to produce variations in the geometric structure.

We employed a technique to transfer learning experience in training. Firstly, the model was pre-trained on Microsoft Common Objects in Context (MS COCO) dataset (Lin et al., 2014). Consequently, an output layer of the model was initialized according to EAD class numbers. Lastly, the model was re-trained with EAD data. Regardless of the large-scale domain gap between natural and medical images, the underlying representational features continued to share certain commonalities. Due to the small volume of the EAD data, our transfer strategy had a considerable advantage than the "train-from-scratch" method when training on EAD dataset.

In this task, we explore a variety of training strategies to resolve the multi-label classification problem at pixel level. Some of the pixels in an image may be associated with multiple labels. Therefore, it is not appropriate to use classical softmax with cross-entropy loss function (L) to resolve this problem. Denote the finite set of labels as: $L=\{l_0, l_1, \ldots, l_K; l_k\}$; $M=\{l_i, \ldots, l_j; i, j=0, 1, \ldots, K$ and $i\leqslant j\}$ is a subset of L, where K is the number of classes in L. Dataset consisting of N training samples is denoted as $\{(x_0, M_0), (x_1, M_1), \ldots, (x_n, M_n); n=1, 2, \ldots, N\}$, where x denotes training samples, and then the objective function can be expressed as follows:

$$\min_{\theta} \sum_{n=0}^{N} L(f(x_n; \theta), M_n),$$

where f is the network and θ represents the model parameters that need to be optimized. We can decompose this multi-label classification problem into a set of binary classification tasks $\{\tau_0, \tau_1, \ldots, \tau_k; k=0, 1, \ldots, K\}$. We can resolve each sub-task independently if there is no correlation between τ_k. However, this assumption is not valid considering the task. Essentially, the probability that a certain pixel is classified into classes i or j may not be equal, but there is a certain degree of co-occurrence.

To model the more flexible co-occurrence relationship, we constructed a CC method: each binary classifier B_k uses the image and prediction results of all previous stages as input to produce results of the current stage. Characteristically, the predicted results of ith stage O_i can be expressed as:

$$O_i \leftarrow B_k(x, (O_0, O_1, \ldots, O_{i-1})),$$

where $O_0, O_1, \ldots, O_{i-1}$ represent the outputs of the previous $k-1$ classifiers. The order of classifiers is fixed, related to the segmentation difficulty in each

category (order from easy to difficult is medical devices, over-exposed areas, bubbles, specular reflection, and other artifacts).

In terms of network architecture, we investigated three mainstream segmentation structures: UNet (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), and DeepLabv3 (Chen et al., 2018). We noted that PSPNet achieved relatively optimal performance.

EAD2019 included three subsets, each corresponding to three tasks: artifact detection, semantic region segmentation, and detection generalization. Regarding the artifact detection and generalization tasks, the training set contained 2613 images. We randomly selected 2322 images for training and 291 images for validation. K-fold cross validation was not considered because of the limited data size which may bring performance fluctuation in the process. We will elaborate on the same if more data are collected in a future study. The sample numbers of test sets for detection and generalization tasks were 195 and 51, respectively. In the segmentation task, 474 samples from 589 images were included in the training, and the rest was used as a test set.

We trained on four graphic processing units (GPUs) with a batch-size of 8, and augmented training data by using random scaling, random flipping, and color normalization. Stochastic gradient descent algorithm with momentum was used in the optimization process; the initial learning rate was set to 0.02 with a scheduling strategy of multistep. Cascaded R-CNN had three stages, and each stage was trained in sequence. The thresholds of RPN were set to 0.5, 0.6, and 0.7 for each training session, respectively. Table 1 lists the improvements in the performance using different techniques sequentially. Notably, data augmentation effectively prevents a model from overfitting, while pre-training further enhances the generalization of the final model.

Table 1 Performance gains on detection and generalization tasks using different techniques

Technique	Score_d (detection)	Score_g (generalization)
Baseline	0.1800	
+Data augmentation	0.3105	0.3600
+Validation set	0.3215	0.3600
+COCO pre-training	0.3395	0.3900
+Adjust NMS parameters	0.3429	0.3500

NMS: non-maximum suppression

To illustrate the effectiveness of our cascade detection method, we compared the detection performance of networks with/without the cascade structure. The results demonstrated that the cascade structure could significantly improve the detection score of different types of artifacts (Table 2).

Table 2 Comparison of detection performance of different network architectures

Model	Score_d
YOLOv3	0.2200
Cascade R-CNN	0.3429

Fig. 5 shows the results of our proposed method on EAD dataset qualitatively. The size variances of different classes are significantly large. This also reflects on the specialization of medical image detection, which differs from general object detection.

We evaluated several backbone networks for the segmentation task, including UNet with VGG-11 and VGG-16, PSPNet with residual structure, and DeepLabv3 with residual structure. PSPNet with a structure of ResNet-34 achieved the best results.

We divided the entire dataset into training and test sets according to a ratio of 4:1. To perform data augmentation, the original image was first resized into five different spatial scales (0.7, 0.9, 1.0, 1.2, and 1.5 times of the original size); then, the scaled images were randomly cropped to 224×224. Next, the processed images were randomly flipped horizontally or vertically and rotated at a certain angle. Thereafter, the augmented images were transferred to a CC for training, in which each classifier was initialized with pre-trained weights. Finally, the predicted results at current resolution were mapped back to the original image by interpolation from the nearest neighbor.

Similar to the detection task, we used a batch size of 8 to train the four GPUs. Stochastic gradient descent algorithm with momentum was used for optimizing, where the weight decay was set to 5×10^{-4}, and the initial learning rate was 0.01. In the first 100 epochs, five epochs were deemed as a cycle to adjust the learning rate cyclically.

Table 3 compares the performance between DeepLabv3 and PSPNet. DeepLabv3 is trained with five independent binary classifiers, while PSPNet corresponds to the CC structure.

Fig. 6 shows some results of our proposed method on segmentation set. We observe that the model learns the apparent characteristics of various artifact classes well, and can outline the region contours accurately.

Table 3 Performance comparison between DeepLabv3 and PSPNet

Model	Overlap	F2 score	Score_s
DeepLabv3	0.5775	0.6136	0.5865
PSPNet	0.5986	0.6225	0.6046

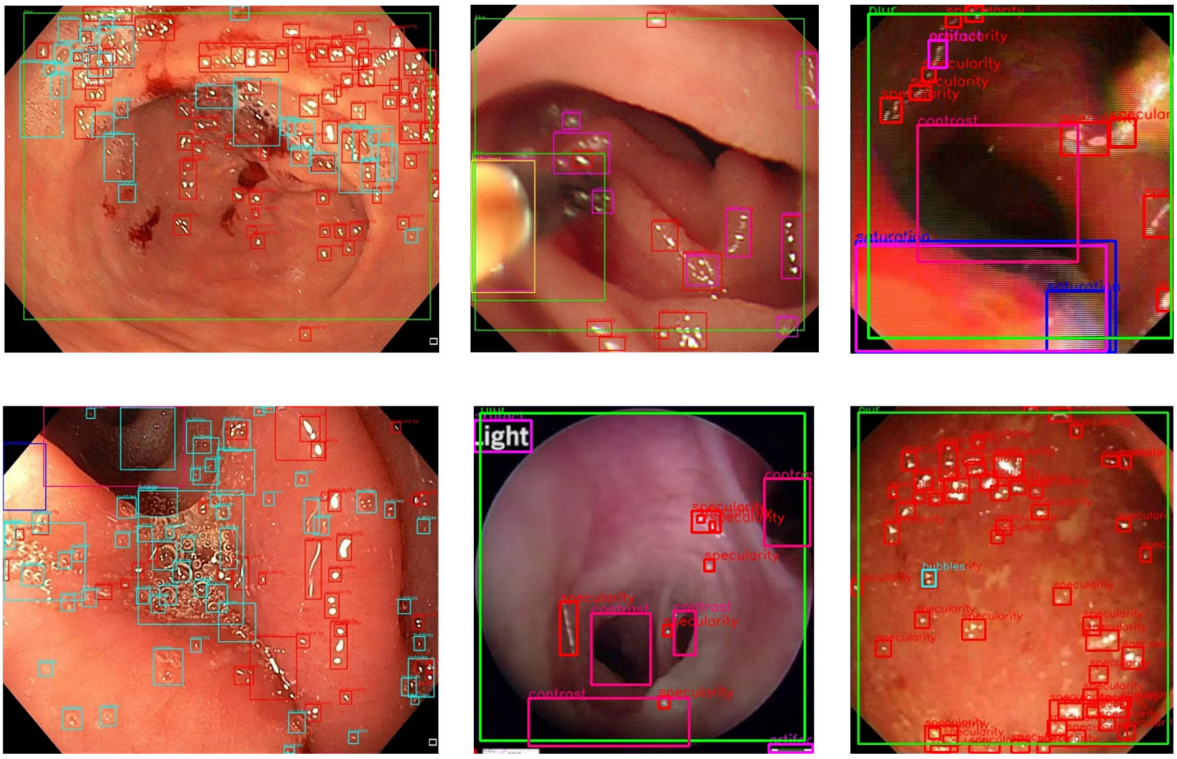

Fig. 5 Qualitative detection results of method on EAD dataset

The red boxes represent specular reflection, the light blue boxes represent bubbles, the dark blue boxes represent supersaturated regions, the green boxes represent blurs, and the purple red boxes represent other defects

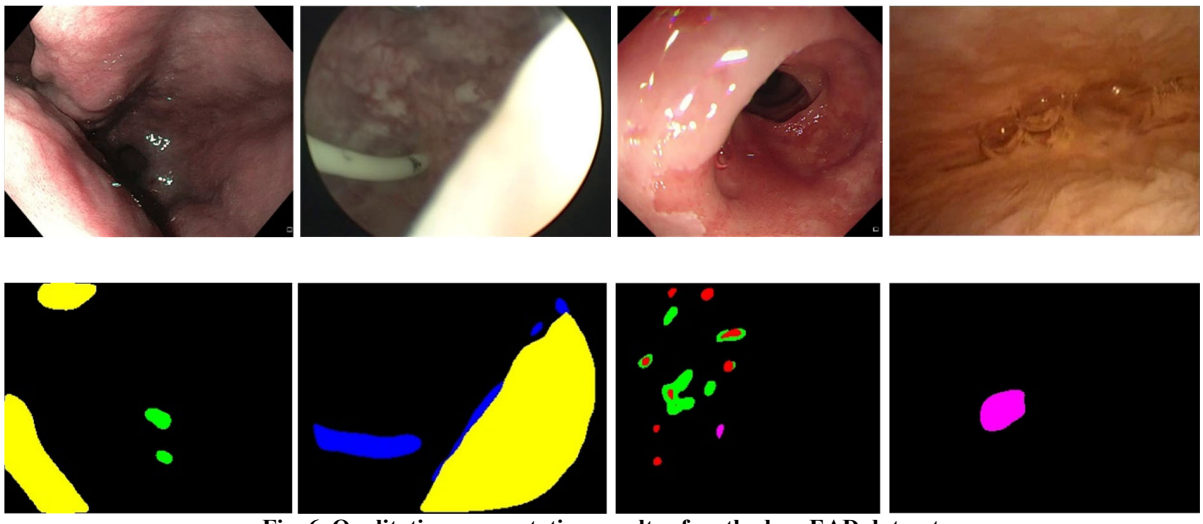

Fig. 6 Qualitative segmentation results of method on EAD dataset

The yellow region represents the supersaturated region, the blue region represents the artificial devices, the green region represents the specular reflection, the purple region represents the bubble, and the red region represents other artifacts

In this paper, we propose two methods, based on deep neural network, to perform the multi-class artifacts detection and segmentation in endoscopic video imaging environment. For the detection task, we integrate the FPN with neural network, and introduce the idea of cascade detection. For the segmentation task, we revise the classical PSPNet network and combine it with CC. We have achieved inspiring results on both tasks. This technology is expected to be applied to automated or semi-automated endoscopic organ examination in the near future, so as to improve the efficiency of medical workers.

Contributors

Yan-yi ZHANG wrote the manuscript after integrative data analyses. Di XIE led the methodology and comparative experiments. Both authors read and approved the final manuscript and, therefore, had full access to all the data in the study and take responsibility for the integrity and security of the data.

Acknowledgments

We thank Tao SONG, Hai-ming SUN, Qi YAO, Liang-jun ZHANG, and Jian ZHOU (Hikvision Research Institute, Hangzhou, China) for helpful discussion and insightful advices for this paper.

Compliance with ethics guidelines

Yan-yi ZHANG and Di XIE declare that they have no conflict of interest.

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Informed consent was obtained from all patients for being included in the study. Additional informed consent was obtained from all patients for whom identifying information is included in this article.

References

Ali S, Zhou F, Daul C, et al., 2019. Endoscopy artifact detection (EAD 2019) challenge dataset. arXiv preprint, arXiv: 1905.03209. https://doi.org/10.17632/c7fjbxcgj9.1

Allan M, Shvets A, Kurmann T, et al., 2019. 2017 Robotic instrument segmentation challenge. arXiv preprint, arXiv: 1902.06426.

Bouget D, Benenson R, Omran M, et al., 2015. Detecting surgical tools by modelling local appearance and global shape. *IEEE Transact Med Imag*, 34(12):2603-2617. https://doi.org/10.1109/TMI.2015.2450831

Bouget D, Allan M, Stoyanov D, et al., 2017. Vision-based and marker-less surgical tool detection and tracking: a review of the literature. *Med Image Anal*, 35:633-654. https://doi.org/10.1016/j.media.2016.09.003

Cai ZW, Vasconcelos N, 2018. Cascade R-CNN: delving into high quality object detection. IEEE/CVF Conference on Computer Vision and Pattern Recognition, p.6154-6162. https://doi.org/10.1109/CVPR.2018.00644

Chen LC, Zhu YK, Papandreou G, et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European Conference on Computer Vision, p.833-851. https://doi.org/10.1007/978-3-030-01234-2_49

Kalalembang E, Usman K, Gunawan IP, 2009. DCT-based local motion blur detection. International Conference on Instrumentation, Communication, Information Technology, and Biomedical Engineering, p.1-6. https://doi.org/10.1109/icici-bme.2009.5417252

Leibetseder A, Primus MJ, Petscharnig S, et al., 2017. Real-time image-based smoke detection in endoscopic videos. Proceedings of the on Thematic Workshops of ACM Multimedia, p.296-304. https://doi.org/10.1145/3126686.3126690

Lin TY, Maire M, Belongie S, et al., 2014. Microsoft COCO: common objects in context. *In*: Fleet D, Pajdla T, Schiele B, et al. (Eds.), Computer Vision—ECCV 2014. Springer, p.740-755. https://doi.org/10.1007/978-3-319-10602-1_48

Loukas C, Georgiou E, 2015. Smoke detection in endoscopic surgery videos: a first step towards retrieval of semantic events. *Int J Med Robot Comput Assist Surg*, 11(1):80-94. https://doi.org/10.1002/rcs.1578

Münzer B, Schoeffmann K, Böszörmenyi L, 2013. Detection of circular content area in endoscopic videos. Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, p.534-536. https://doi.org/10.1109/CBMS.2013.6627865

Münzer B, Schoeffmann K, Böszörmenyi L, 2018. Content-based processing and analysis of endoscopic images and videos: a survey. *Multimed Tools Appl*, 77(1):1323-1362. https://doi.org/10.1007/s11042-016-4219-z

Read J, Pfahringer B, Holmes G, et al., 2011. Classifier chains for multi-label classification. *Mach Learn*, 85(3):333-359. https://doi.org/10.1007/s10994-011-5256-5

Ren SQ, He KM, Girshick R, et al., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems, p.91-99. https://doi.org/10.1109/tpami.2016.2577031

Ronneberger O, Fischer P, Brox T, 2015. U-Net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-assisted Intervention, p.234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Shvets AA, Rakhlin A, Kalinin AA, et al., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning. 17th IEEE International Conference on Machine Learning and Applications, p.624-628. https://doi.org/10.1109/ICMLA.2018.00100

Zhao HS, Shi JP, Qi XJ, et al., 2017. Pyramid scene parsing

network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p.6230-6239. https://doi.org/10.1109/cvpr.2017.660

中文概要

题　目：内窥镜视频中多类别人造物的检测与分割

概　要：为准确定位内窥镜视频中的人造物，帮助医生提升诊断准确率，引入深度神经网络检测与分割模型，采用特征金字塔与级联 R-CNN 相互结合的框架，并使用 PSPNet 结合分类器链的思想，从而解决分割及数据匮乏问题，有效提升性能，并在 EAD 2019 数据集上取得领先的性能。

关键词：内窥镜；检测与分割；多类别人造物；级联 R-CNN；PSPNet