Upper gastrointestinal anatomy detection with multi-task convolutional neural networks

Zhang Xu[1], *Yu Tao*[1], *Zheng Wenfang*[2,3], *Lin Ne*[2,3], *Huang Zhengxing*[1], *Liu Jiquan*[1] ✉, *Hu Weiling*[2,3], *Duan Huilong*[1], *Si Jianmin*[2,3]

[1]*Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou 310027, People's Republic of China*
[2]*Department of Gastroenterology, Sir Run Run Shaw Hospital, Medical School, Zhejiang University, Hangzhou, 310016, People's Republic of China*
[3]*Institute of Gastroenterology, Zhejiang University, Hangzhou 310029, People's Republic of China*
✉ *E-mail: liujq@zju.edu.cn*

Published in Healthcare Technology Letters; Received on 16th September 2019; Accepted on 2nd October 2019

Esophagogastroduodenoscopy (EGD) has been widely applied for gastrointestinal (GI) examinations. However, there is a lack of mature technology to evaluate the quality of the EGD inspection process. In this Letter, the authors design a multi-task anatomy detection convolutional neural network (MT-AD-CNN) to evaluate the EGD inspection quality by combining the detection task of the upper digestive tract with ten anatomical structures and the classification task of informative video frames. The authors' model is able to eliminate non-informative frames of the gastroscopic videos and detect the anatomies in real time. Specifically, a sub-branch is added to the detection network to classify NBI images, informative and non-informative images. By doing so, the detected box will be only displayed on the informative frames, which can reduce the false-positive rate. They can determine the video frames on which each anatomical location is effectively examined, so that they can analyse the diagnosis quality. Their method reaches the performance of 93.74% mean average precision for the detection task and 98.77% accuracy for the classification task. Their model can reflect the detailed circumstance of the gastroscopy examination process, which shows application potential in improving the quality of examinations.

1. Introduction: Since gastroscopy is able to observe the interior of the gastrointestinal (GI) tract directly, it has been widely used for GI examinations [1]. The conventional gastroscopy procedure usually needs the clinicians to comprehensively observe the anatomies of the upper digestive tract of the patients, including oesophagus, stomach and duodenum, to make a diagnosis. Due to operating habits, many doctors tend to miss some mucosal areas or fail to observe some physiological, anatomical structures during examinations, resulting in missed diagnosis [2]. Furthermore, for inexperienced doctors, they may tend to over-close the lens to gastric mucosa, which may cause a large number of non-informative frames and affects the quality of the gastroscopy. However, there is currently no effective method to supervise and evaluate the quality of the gastroscopy, and there is no quantitative assessment to determine whether the doctor has performed an efficient examination of each anatomy.

Recently, deep learning, especially the convolutional neural network (CNN) has been applied in the medical domain [3] and has demonstrated success in diverse medical image analysis tasks [4, 5]. Endoscopy is one of them [6, 7], including the classification and detection of precancerous diseases [8, 9] and early gastric cancer screening under conventional endoscopy [10, 11]. When we find or depict a lesion, it is necessary to identify the anatomical locations. However, most researchers focus on using computer vision techniques to detect lesions under endoscopy. Few studies focus on the recognition of the anatomies, which is the first crucial step for GI diseases. Without learning these normal anatomical features, it is very difficult to recognise abnormalities and to diagnose diseases properly. In [12], a CNN-based diagnostic program was constructed to classify four major anatomical locations (larynx, oesophagus, stomach and duodenum) and three subsequent sub-classifications. However, the division of the stomach is relatively rough, which may not suitable for real clinical situations. In [13], researchers tried to train the deep CNN to classify gastric locations into 10 or 26 parts, with the accuracy of 90 and 65.9%, respectively. However, the accuracy of 26 parts is not outstanding and classifying anatomies cannot ensure the lesions' location.

To solve the problems mentioned above, we propose a real-time multi-task anatomy detection network (MT-AD) based on CNN for gastroscopy inspection. MT-AD can perform two tasks simultaneously. (i) Classify three image types, including informative and non-informative images under white-light gastroscopy and narrow-band imaging (NBI) [14]. Non-informative images are images with a large area of specific artefacts including motion blur, defocus, specular reflections and bubbles. (ii) Detect ten gastric anatomies in real time, including oesophagus, dentate line, cardia (observed from the inside of the stomach), fundus, body, antrum, angle, pylorus, duodenal bulb (DB) and duodenal descending part (DDP). The specific workflow of MT-AD is shown in Fig. 1. When a video is fed into MT-AD, it performs one forward calculation for each frame. If the classification result shows that the frame is not informative, the detection results will not be displayed on the related frame. Consequently, we can count the number of frames of specific anatomy appearing in an examination, and thus the precise time at which each anatomy is effectively examined can be determined. This will help reflect the quality of the gastroscopy inspection process and assist the clinicians in screening the areas that are easily missed during the examinations, and therefore decreasing the rate of missed diagnosis.

The main contributions of this Letter are as follows:

(i) We construct a CNN-based gastric anatomy detection network called MT-AD for evaluating the quality of gastroscopy for the first time. By detecting anatomies, it is also promising to help doctors understand the lesions further.
(ii) The MT-AD model can not only detect the anatomies, but also identify whether the target image is informative.

doi: 10.1049/htl.2019.0066

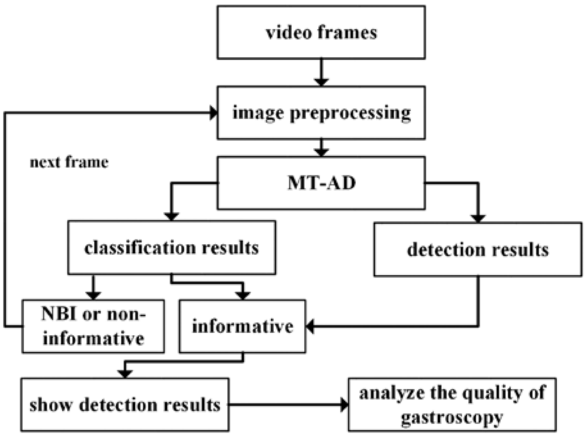

Fig. 1 *Workflow of the proposed method*

(iii) We figure out the ratio of the informative frames to the non-informative frames in a gastric video and we obtain the number of informative frames that each anatomy is checked. These indicators can reflect the quality of gastric videos and show the potential to help the clinician improve operating gastroscopy skills.

2. Related works

2.1. Object detection networks: Object detection involves locating and classifying objects. In the deep learning era, powered by CNN, object detection approaches can be roughly categorised into two main types of pipelines, namely, two-stage approaches [15, 16] and one-stage approaches [17]. Two-stage approaches divide the object detection task into two stages: extract regions of Interest (ROI), then classify and regress ROI. One-stage approaches remove ROI extraction process and directly classify and regress the candidate anchor boxes, including Single Shot MultiBox Detector (SSD) [18], which uses features from different convolutional layers to regress and classify the anchor boxes to get high performance. Although one-stage approaches perform not as good as two-stage approaches on small object detection, they have advantages on-time performance.

As shown in Fig. 2, most of the classes belong to large objects in pixel level. We finally choose SSD as the network for the detection task in MT-AD. Although there are many detection networks, which can achieve higher precisions, they usually use more complex modules [19], deepen or widen the networks [20], increase the models' parameters [21], increase the input size [22] and sacrifice time performance [19]. However, we aim to construct a real-time detection network and time performance is paramount.

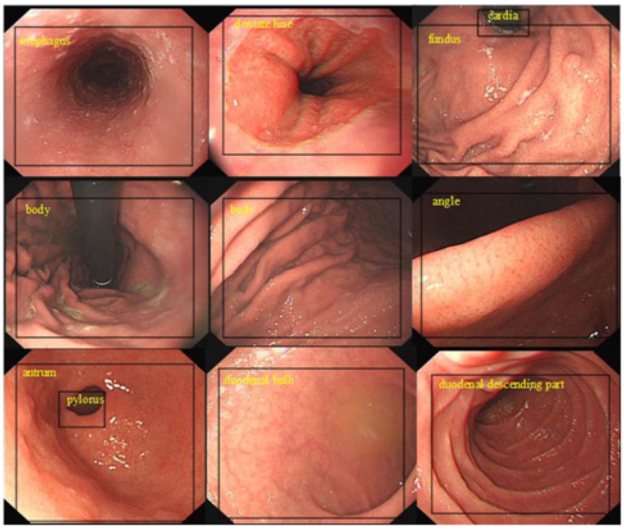

Fig. 2 *Examples of the labelled images for each class*

2.2. Multi-task learning (MTL) networks: MTL is a form of inductive transfer. The inductive transfer can help improve a model by introducing an inductive bias, which causes a model to prefer some hypotheses over others. MTL usually shares the same backbone of convolutional layers, while learning task-specific layers [23].

Suppose we have two related tasks A and B to solve. The conventional approach is to train a model for each task (see Fig. 3 'Solution 1'), while MTL enables the tasks to share the same backbone (see Fig. 3 'Solution 2'). Compared with the conventional approach, MTL can reduce model complexity, decrease the computation and improve time performance [24]. However, conventional MTL networks usually combine multiple 'losses' together for optimisation. Our MT-AD networks only introduce

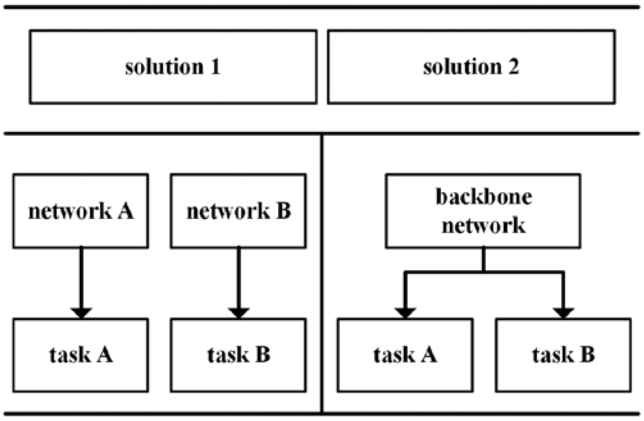

Fig. 3 *Difference between MTL and the conventional approach*

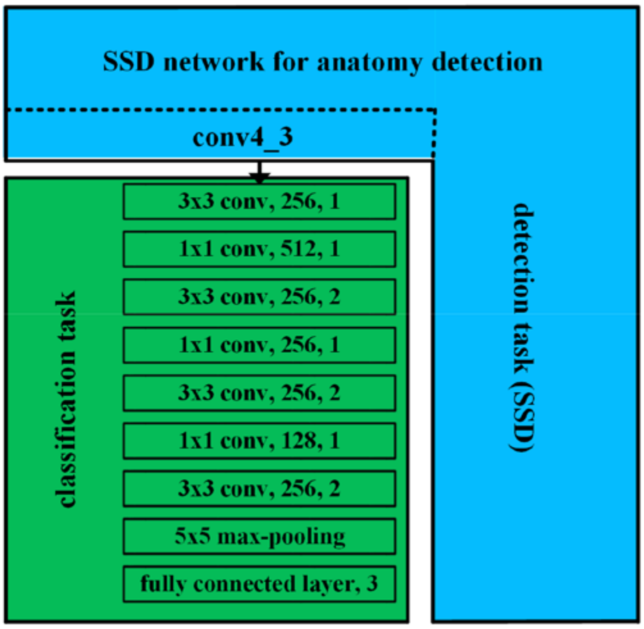

Fig. 4 *Architecture of the MT-AD*

doi: 10.1049/htl.2019.0066

a similar architecture from MTL. The training strategies are essentially different from each other (see Section 3.2).

3. Methods: We design a multi-task neural network called MT-AD to realise gastric image classification and anatomy detection at the same time (see Fig. 4). The blue part depicts detection architecture, which is introduced from SSD [18] without any modification. The green part depicts the classification architecture to classify NBI images, informative images and non-informative images under conventional endoscopy. We set the output of 'Conv4_3 layer', which is one of the SSD layers, as the input of classification architecture.

3.1. Classification architecture: Fig. 4 shows the classification network. There are seven convolutional layers, one max-pooling layer and one fully connected layer. For convolutional layers, the parameters include the kernel size, the output feature channel and the stride. The kernel size of the max-pooling layer is set as 5 × 5. The output channel of the fully connected layer is set as 3, which is the same as the classes of the classification task. It shares the same backbone with the detection task. The input of the classification network is the output of the Conv4_3, which is one of the SSD layers.

3.2. Training strategies: Since the detection task is more complex than the classification task, we first fix the parameters of the classification network to train the SSD network. After that, we fix the parameters of SSD (including the backbone network parameters) to train the classification network. The proposed networks were implemented using Python on the Pytorch [25] package based on the GeForce RTX 2080Ti.

3.2.1 Training SSD: We apply the pre-trained model, which was downloaded from SSD's GitHub (https://github.com/weiliu89/caffe/tree/ssd) to initialise model parameters. The input image size of the proposed network is 300 × 300. The batch size is set as 8. The basic learning rate is 0.0005. The training undergoes 120,000 iterations and the learning rate decays after 80,000 and 100,000 iterations. Other hyper-parameters keep the same as conventional SSD.

3.2.2 Training classification network: We randomly initialise the parameters of the classification network. The input image size of the classification network is also 300 × 300. The batch size is set as 8. The basic learning rate is 0.001. The training undergoes 25 epochs. The learning rate decays after 10 and 20 epochs. The momentum is 0.9 and the weight decay is 0.0001. we adopt cross entropy to compute training loss and stochastic gradient descent to update model parameters.

4. Experiments and results

4.1. Dataset preparing: We invited two experienced endoscopists to label 60,233 gastric images (see the examples of the labelled images in Fig. 2) for the detection task and 40,145 gastric images for the classification task. The datasets for detection and classification are independent and do not overlap with each other. For detection dataset, the training images were collected from 1 May to 15 June 2015. The testing images were collected from 16 to 30 June 2015. The details of the datasets are shown in Tables 1–3. All the images were taken from OLYMPUS EVIS LUCERA ELITE CLV-290SL or OLYMPUS EVIS LUCERA ELITE CLV-260SL, which is the model name of the endoscope device from the same hospital.

4.2. Performance of the detection network: We adopt the PASCAL VOC metric of 2010 as the evaluation method. The mean Average Precision (mAP) of the ten anatomies' detection is 93.74% when we set the confidence threshold as 0.5. The precision–recall (*P*–*R*) curve is shown in Fig. 5.

Table 1 Dataset of the detection task

	Training	Testing
patients	2932	843
images	47,623	12,600
labelled boxes	59,513	15,762

Table 2 Images for each class of the detection dataset

Class	Training	Testing
oesophagus	6851	1766
dentate line	4951	1356
cardia	3955	1054
fundus	5684	1503
body	10,723	2932
antrum	7209	1916
angle	5050	1357
pylorus	5930	1597
DB	4832	1175
DDP	4328	1106
total	59,513	15,762

Table 3 Dataset of the classification task

	Training	Testing
informative images	10,053	2000
non-informative images	10,138	2000
NBI images	13,954	2000
total images	34,145	6000

From Fig. 5 and Table 4, we can find that only the average precision (AP) of the dentate line is 84.93%; other anatomies' APs are higher than 90%, which means that our MT-AD performs excellently on detecting anatomies.

4.3. Performance of classification network: The classification task aims to classify informative images, NBI images, and non-informative images. The confusion matrix of the classification results is shown in Fig. 6. We can figure out that the accuracy of

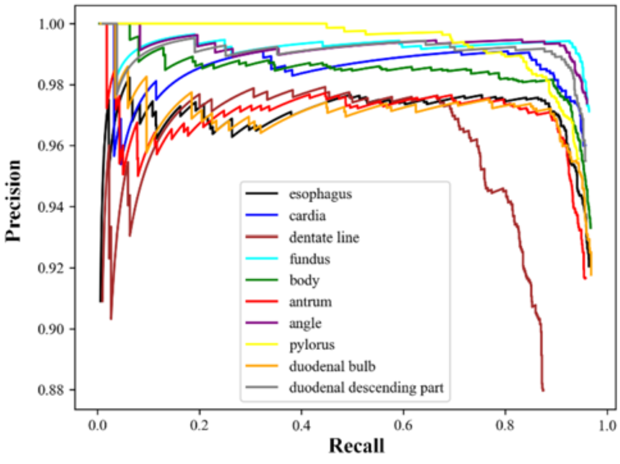

Fig. 5 *P–R curves of the detection task*

Table 4 Average precision of each anatomy

Method	mAP	Oesophagus	Cardia	Dentate line	Fundus	Body	Antrum	Angle	Pylorus	DB	DDP
MT-AD, %	93.74	94.04	94.76	84.93	95.95	95.39	93.38	95.52	93.92	94.42	95.12

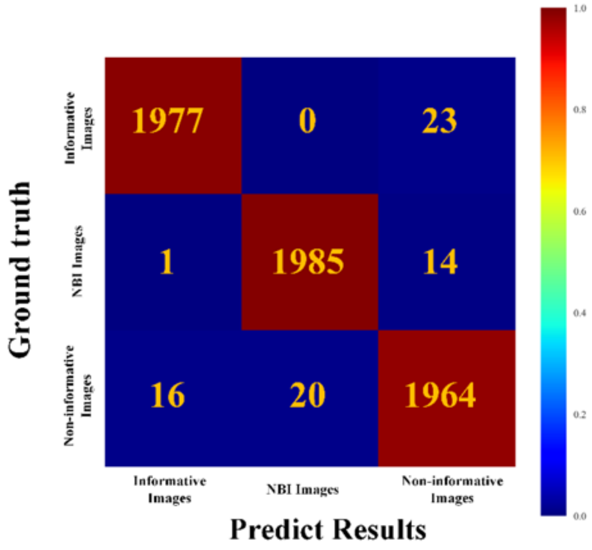

Fig. 6 *Heatmap of confusion matrix from the classification task*

the classification is 98.77%. Each class's F1 score is 99.01, 99.13 and 98.18%. The high accuracy of the classification task is a prerequisite for the high reliability of the final video statistical results of MT-AD.

4.4. Video statistical results from MT-AD: We totally collected 83 gastric videos from a hospital. All the videos which were taken from OLYMPUS EVIS LUCERA ELITE CLV-290SL with 25 FPS were operated by the same doctor. MT-AD can detect these videos in real time with 30.8 FPS, which is promising for clinical application. Fig. 7 shows examples of MT-AD's performance. The first row denotes the input images. The second row shows the results generated by MT-AD. And the classification results determine whether display detection results in the images. This method can reduce false positives to some extent.

Fig. 8 shows the statistical results processed by MT-AD. We can find out that the ratio of informative frames and non-informative frames is about 7:3. From Fig. 9, we can also figure out that the doctor takes more time to check the body of the stomach. On the contrary, the time for checking the cardia and the dentate line is relatively short.

All the indicators mentioned in this Letter can reflect the detailed circumstance of the gastroscopy examination process to some degree. These indicators prove that our model has great application potential value for improving the quality of examinations.

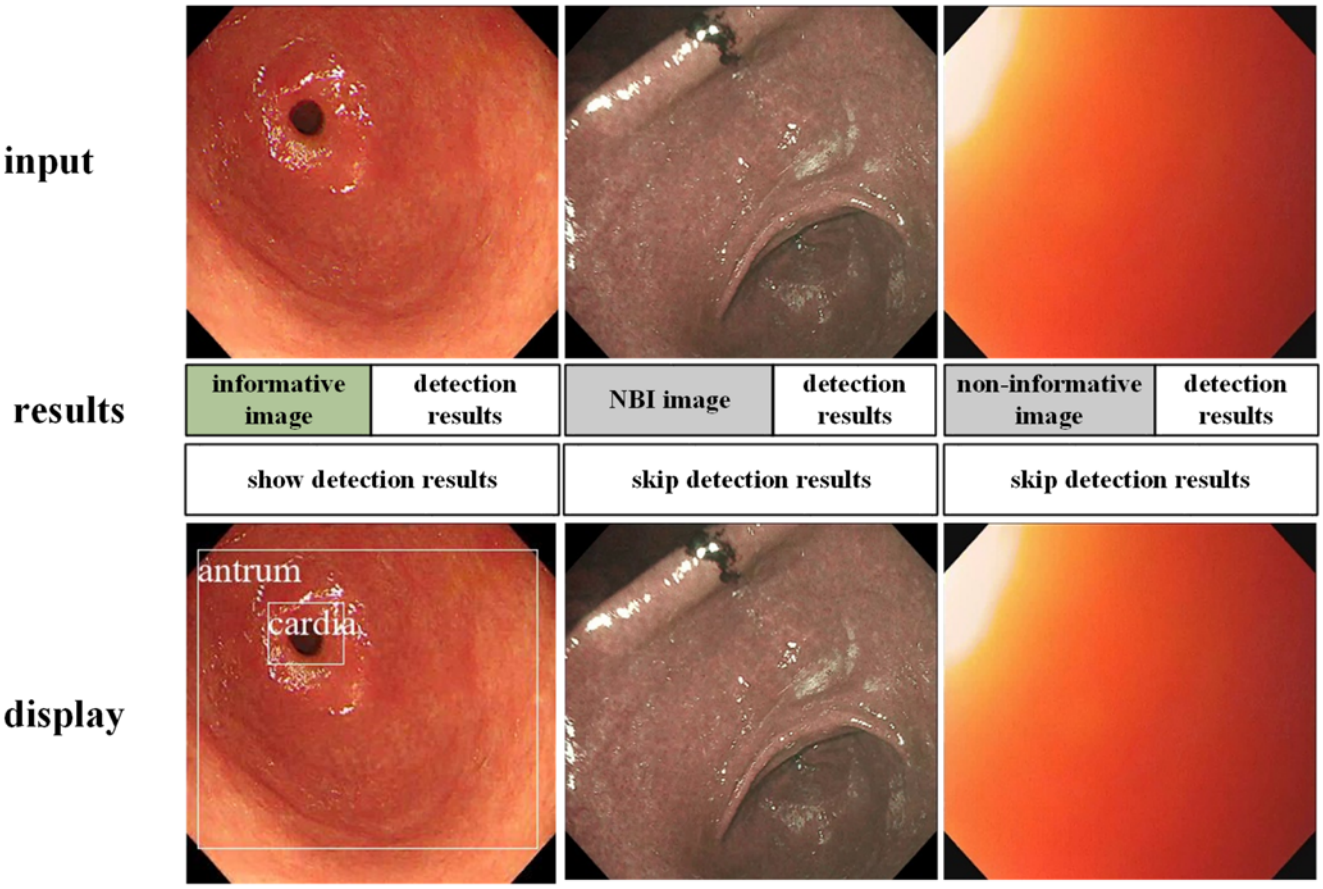

Fig. 7 *Examples of MT-AD's detection results*

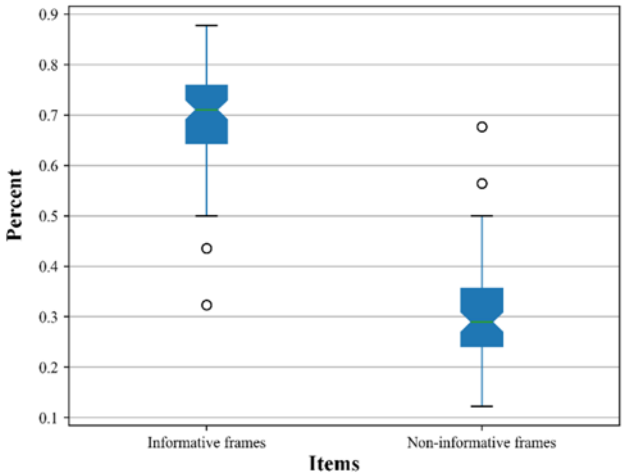

Fig. 8 *Statistical results from 83 gastric videos. Y-axis denotes the ratio of the valid time to the total time of the video (exclude NBI)*

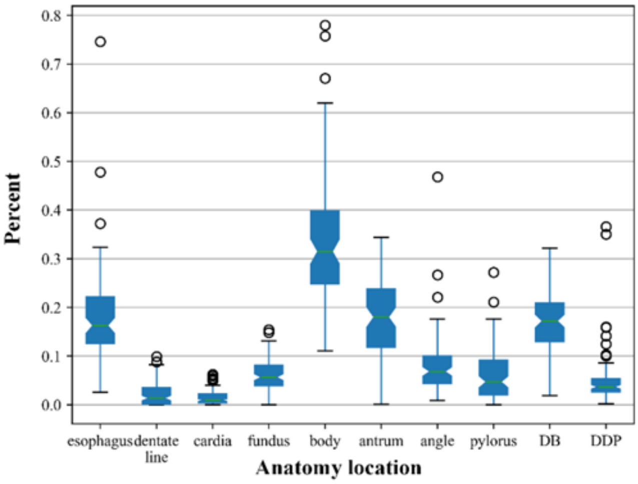

Fig. 9 *Statistical results from 83 gastric videos. Y-axis denotes the ratio of the time at which the specific anatomy is examined to the valid time of this video*

5. Conclusion: In this Letter, we constructed MT-AD based on CNN. Our network can recognise informative frames of the gastroscopic videos and detect the anatomies in real time. The model reaches the performance of 93.74% mAP for the detection task and 98.77% accuracy for classification tasks.

Furthermore, we performed a statistical analysis of the classification and detection results from MT-AD to do a rough evaluation of the inspection quality by using indictors including the ratio of the valid time to the total gastroscopy time (exclude NBI), the ratio of the time at which the specific anatomy is examined to the valid time of this video. The proposed model shows application potential to help the doctor improve the quality of gastroscopy.

However, there are some problems need to be solved in the future. First, the proposed model is not capable of detecting gastric anatomy in narrow-band images. In the future, we will make up for this deficiency. Second, we only give some descriptive statistics on analysing the gastric videos in this Letter. In the future, we will be committed to concluding the reasonable inspection time for each anatomy with clinicians. Third, we will combine the detection of diseases and anatomies in gastroscopy for automatic generation of structured endoscopy examination reports, which will reduce the burdens on doctors.

6. Funding and declaration of interests: This work was supported by the National Key Research and Development Program of China (grant 2017YFC0114106), National Natural Science Foundation of China (grant nos. 31771072 and 81827804).

7 References

[1] Yao K.: 'The endoscopic diagnosis of early gastric cancer', *Ann. Gastroenterol.*, 2012, **26**, (1), pp. 1–11

[2] Zhu R., Zhang R., Xue D.: 'Lesion detection of endoscopy images based on convolutional neural network features'. 2015 8th Int. Congress on Image and Signal Processing (CISP), Shenyang, China, 2015, pp. 372–376

[3] Shen D., Wu G., Suk H.-I.: 'Deep learning in medical image analysis', *Annu. Rev. Biomed. Eng.*, 2017, **19**, (1), pp. 221–248

[4] Anthimopoulos M., Christodoulidis S., Ebner L., ET AL.: 'Lung pattern classification for interstitial lung diseases using a deep convolutional neural network', *IEEE Trans. Med. Imaging*, 2016, **35**, (5), pp. 1207–1216

[5] Pereira S., Pinto A., Alves V., ET AL.: 'Brain tumor segmentation using convolutional neural networks in MRI images', *IEEE Trans. Med. Imaging*, 2016, **35**, (5), pp. 1240–1251

[6] Min J.K., Kwak M.S., Cha J.M.: 'Overview of deep learning in gastrointestinal endoscopy', *Gut Liver*, 2019, **13**, (4), pp. 1–6

[7] Yang Y.J., Bang C.S.: 'Application of artificial intelligence in gastroenterology', *World J. Gastroenterol.*, 2019, **25**, (14), pp. 1666–1683

[8] Zhang X., Hu W., Chen F., ET AL.: 'Gastric precancerous diseases classification using CNN with a concise model', *PloS One*, 2017, **12**, (9), p. e0185508

[9] Zhang R., Zheng Y., Mak T.W.C., ET AL.: 'Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain', *IEEE J. Biomed. Health Inf.*, 2017, **21**, (1), pp. 41–47

[10] Hirasawa T., Aoyama K., Tanimoto T., ET AL.: 'Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images', *Gastric Cancer*, 2018, **21**, (4), pp. 1–8

[11] Sakai Y., Takemoto S., Hori K., ET AL.: 'Automatic detection of early gastric cancer in endoscopic images using a transferring convolutional neural network'. 2018 40th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 4138–4141

[12] Takiyama H., Ozawa T., Ishihara S., ET AL.: 'Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks', *Sci. Rep.*, 2018, **8**, (1), pp. 7497–7504

[13] Wu L., Zhou W., Wan X., ET AL.: 'A deep neural network improves endoscopic detection of early gastric cancer without blind spots', *Endoscopy*, 2019, **51**, (6), pp. 522–531

[14] Li H., Dai J., Xue H., ET AL.: 'Application of magnifying endoscopy with narrow-band imaging in diagnosing gastric lesions: a prospective study', *Gastrointest. Endosc.*, 2012, **76**, (6), pp. 1124–1132

[15] Girshick R., Donahue J., Darrell T., ET AL.: 'Rich feature hierarchies for accurate object detection and semantic segmentation'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, 2014, pp. 580–587

[16] Ren S., He K., Girshick R., ET AL.: 'Faster R-CNN: towards real-time object detection with region proposal networks', Proc. of the Conf. on Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, Canada, 2015, pp. 91–99

[17] Redmon J., Divvala S., Girshick R., ET AL.: 'You only Look once: unified, real-time object detection'. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, 2016, pp. 779–788

[18] Liu W., Anguelov D., Erhan D., ET AL.: 'SSD: single shot multibox detector'. European Conf. on Computer Vision, Amsterdam, Netherlands, 2016, pp. 21–37

[19] Fu C.-Y., Liu W., Ranga A., ET AL.: 'DSSD: deconvolutional single shot detector', arXiv preprint arXiv:1701.06659, 2017, pp. 1–11

[20] Lin T.-Y., Goyal P., Girshick R., ET AL.: 'Focal loss for dense object detection'. Proc. of the IEEE Int. Conf. on Computer Vision, Venice, Italy, 2017, pp. 2980–2988

[21] Jeong J., Park H., Kwak N.: 'Enhancement of SSD by concatenating feature maps for object detection', arXiv preprint arXiv:1705.09587, 2017

[22] Law H., Deng J.: 'Cornernet: detecting objects as paired keypoints'. Proc. of the European Conf. on Computer Vision (ECCV), Munich, Germany, 2018, pp. 734–750

[23] Ruder S.: 'An overview of multi-task learning in deep neural networks', arXiv e-prints, 2017

[24] Vandenhende S., De Brabandere B., Van Gool L.: 'Branched multi-task networks: deciding what layers to share', arXiv e-prints, 2019

[25] Available at https://github.com/pytorch/pytorch, accessed date 2017

doi: 10.1049/htl.2019.0066