Temporal Context Framework for Endoscopy Artefact Segmentation and Detection

Haili Ye[1,2], Hanpei Miao[1,2], Jiang Liu[1,2], Dahan Wang[3] and Heng Li[1,2]

[1]Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen 518055, China
[2]Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[3]Department of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361004, China

Abstract

Endoscopic video processing could facilitate pre-operative planning, intra-operative image guidance and generation of post-operative analysis of the surgical procedure. However, most of the current methods are still based on a single frame of image analysis, which makes the results of the previous frame images independent of each other and causes vibration. In this paper, we propose an temporal context framework for endoscopy artefact segmentation and detection. The framework extends the general segmentation and detection model to the form based on temporal input, and we add a Temporal Context Transformer(TCT) after the encoder of the model to improve the model's ability to construct temporal context features. the experiments of the EndoCV 2022 challenge dataset that this framework can improve the robustness of the model.

Keywords

Medical Image Analysis, Colonoscopic Image, Semantic Segmentation, Object Detection

1. Introduction

Colon cancer[1] is a common malignant tumor of the digestive tract that occurs in the colon. Colon cancer is closely related to the consumption of red meat (such as beef). Incidence of gastrointestinal tumors accounted for the third place. Colon cancer is mainly adenocarcinoma, mucinous adenocarcinoma, undifferentiated carcinoma. Endoscopy[2] can clearly find intestinal lesions, but also can treat some intestinal lesions, such as: intestinal polyps and other benign lesions under the microscope directly removed, intestinal bleeding under the microscope to stop bleeding, the removal of foreign bodies in the colon. Endoscopic video[3] processing could facilitate pre-operative planning, intra-operative image guidance and generation of post-operative analysis of the surgical procedure. Computer assisted interventions[4] have the potential to enhance the surgeon's visualization and navigation capabilities and postoperative analytics to provide insights for surgical training and risk assessment. A necessary element for these processes is scene understanding and, in particular, anatomy and instrument detection and localization. Therefore, by segmenting and differentiating among the elements that appear in the Endoscopic view, it is possible to assess tissue-instrument interactions and understand endoscopic workflow.

Semantic segmentation[5] and object detection[6] are two hot research fields in computer vision. In medical semantic segmentation, Olaf et al proposed a classic medical image segmentation model U-Net[7], and the relevant encoder-decoder structure and skip-layer connection method have great inspiration for subsequent research work. On this basis, a series of novel and effective models are developed, such as U-Net++[8], nnUNet[9], DANet[10], Deeplab[11] and so on. For the analysis of endoscope images, The PraNet[12] proposed by Fan et al. aggregates features at a high level through the parallel partial decoder (PDD) to obtain context information and generate a global map. In medical object detection, Ross et al. proposed the Faster RCNN[13] achieves end-to-end object detection based on a deep learning two-stage structure. Cai et al. proposed Cascade R-CNN[14] to continuously optimize the prediction results by cascading several detection networks. The Swin Transfromer[15] proposed by Liu et al. is a general vision structure designed based on the concept of Transfromer[16], which has achieved breakthroughs in multiple vision tasks. However, most of the current methods are still based on single-frame image analysis, which makes the analysis results not well combined with temporal context information.

Endoscope image sequence can provide more information than single frame image [17, 18], and combining the contextual time information of the before and after images can effectively improve the analysis performance of endoscopy artefact. Inspired by this, in this paper, We propose a temporal Context Framework for endoscopy artefact segmentation and detection. Our contributions are as follows:

- We introduce a general framework to extract temporal context features from sequential images and

4th International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV2022) in conjunction with the 19th IEEE International Symposium on Biomedical Imaging ISBI2022, March 28th, 2022, IC Royal Bengal, Kolkata, India

yehl@mail.sustech.edu.cn (H. Ye)

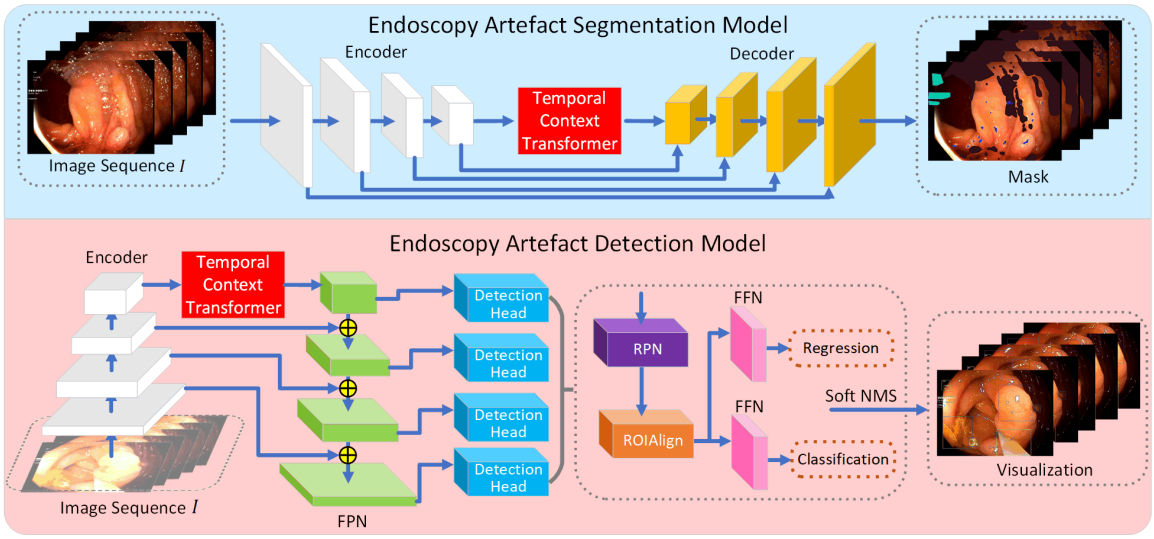

Figure 1: Overall Temporal Context Framework for Endoscopy Artefact Segmentation and Detection

apply it to semantic segmentation and object detection.

- In order to improve the feature modeling ability of the framework,we designed a Temporal Context Transformer(TCT) to improve the feature extraction ability of temporal context.
- Our framework can be adapted to various types of backbone models and can be extended for similar endoscopic analysis problems.

2. METHODOLOGY

In this section, we introduce the proposed temporal context framework for endoscopy artefact segmentation and detection. The overall of this framework as shown in Fig. 1. The framework includes endoscopy artefact segmentation model and endoscopy artefact detection model. The input of both models is the endoscope image sequence, and we set a hyperparameter L to represent the length of the image sequence, so L-frame sequence of input to the model can be represented as $I \in R^{L,3,H,W}$.

In the endoscopy artefact segmentation model, we use the classical coding-decoding results. In particular, the encoder of the model is similar to the traditional encoder, which is responsible for extracting the features of single frame influence. N group temporal context transformer is connected at the end of the encoder to establish the correlation between the image features of each frame. Compared with general single-frame image-based methods, this module utilizes feature correlations between different frames. This process can repair the wrong features extracted by the model, which effectively improves the robustness of the model. We also refer to UNet's hop connection and connect the corresponding encoder and decoding to supplement the shallow feature. The features are integrated by the temporal context transformer and then enter the decoder to obtain the segmentation mask of the endoscopy image.

The input form of the endoscopy artefact detection model is also L-frame image sequence, and the overall structure is similar to the common two-stage target detection model.The feature of each frame is extracted from the encoder, and then the feature is integrated into the N group temporal context transformer. The network structure of feature pyramid can deal with the multi-scale change problem in object detection with a small amount of computation. So the model have a feature pyramid to improve the model's localization ability for multi-scale surgical examples. The multi-scale features extracted by FPN[19] will be input into the corresponding detection head for prediction. The detection head uses a region proposal network(RPN)[13] to filter out suggestion boxes that may have instances of surgical instruments. And ROI Align[13] adopts the corresponding local features in the global features according to these proposal boxes. These local features provide the FFN to classify the artefact within the proposal box and regress to specific coordinates. Finally, the prediction results of different scales are merged and filtered using Sotf NMS[20]. Sotf NMS will remove the prediction results with large overlap and the prediction box, and retain the results with high confi-

dence. The loss function form of object detection model is the same as that of Faster RCNN[13].

Temporal Context Transformer. For the image sequence, there is a little correlation between the image data of the next frame and the next frame. Especially in the case of blur or artifact in the image, introducing the features of the previous frame can effectively repair the situation of target loss or category recognition error. In order to effectively improve the context understanding and feature integration capabilities of the model for image sequences. We designed the temporal context transformer, as show in Fig .2. Temporal context transformer is divided into transformer encoder and transformer decoder. The features extracted in the encoder will be input to the transformer encoder. For the Transformer encoder of layer n, the input is the output $E_{n-1} \in R^{L,C}$ of the upper layer. The coordination transformer encoder has a similar structure to the traditional Transformer encoder, but the difference is that we design the timing code T combining the characteristics of image sequence. The time difference between the two frames can be calculated in the endoscope image sequence and the time sequence coding between different frames can be modeled by normalization of the time difference. When the image sequence length is L, the sequence encoding T_s is a square matrix of $L \times L$:

$$\overline{T} = \begin{bmatrix} 0 & |t_0 - t_1| & \cdots & |t_0 - t_L| \\ |t_1 - t_0| & 0 & \cdots & |t_1 - t_L| \\ \vdots & \vdots & \ddots & \vdots \\ |t_L - t_0| & |t_L - t_1| & \cdots & 0 \end{bmatrix} \tag{1}$$

$$T_s = Normal(\overline{T}^{-1}) \tag{2}$$

In self-attention generates query $Q \in R^{L,C}$, key $K \in R^{L,C}$, and value $V \in R^{L,C}$ based on E_{n-1}. Then calculate the initial self-attention weight $\overline{A} \in R^{L,L} = Softmax((Q*W_Q^E)*(K*W_K^E)^T/\tau)$ between L frames. Then, sequence coding is introduced to calculate the final self-attention weight $A \in R^{L,L} = \overline{A} * T_s$. In this way, the temporal relevance in the original self-attention weight can be strengthened. The following steps are the same as for a classical transformer[16].

The transformer decoder is responsible for decoding and reconstruction of the features of the transformer encoder. The input form of the layer N transformer encoder is $E_{n-1} \in R^{L,C}$. Like the transformer encoder, sequence coding is added to the transformer decoder to improve the temporal modeling ability of the model. In the transformer decoder, the first step is the mask self-attention, which emphasizes the prediction of the model in accordance with the sequence of images. Different from the classical transformer, we add the cross attention[16] unit at the end of the transformer decoder.The transformer decoder calculates query

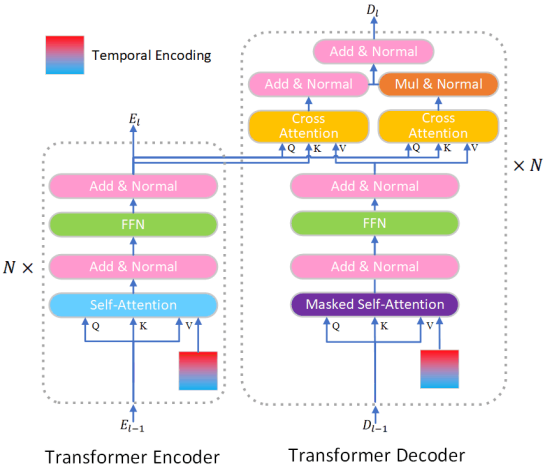

Figure 2: Structure of Temporal Context Transformer(TCT)

$Q \in R^{L,C}$ and key $K \in R^{L,C}$ using the output E_n of the transformer encoder of the same layer. The cross attention weight matrix T_c is calculated by Q and K of transformer encoder. As shown in Fig.2, there are two parallel attention modules for feature learning in this part. We hope that these two attention modules can learn feature compensation and contraction respectively. Therefore, the parameters of the two modules do not share, and matrix addition and matrix cross product are used respectively.The specific operations are as follows:

$$C^{'} = Softmax((Q * W_{Q_1}^D) * (K * W_{K_1}^D)^T/\tau) \tag{3}$$

$$C^{''} = Softmax((Q * W_{Q_2}^D) * (K * W_{K_2}^D)^T/\tau) \tag{4}$$

$$\begin{aligned} D_n = Ins.Norm\{Ins.Norm\{C^{'} * V * W_{V_1}^D + V\} \\ + Ins.Norm\{(C^{''} * V * W_{V_2}^D) \otimes V\}\} \end{aligned} \tag{5}$$

The above process makes the features of each frame images fully fused, and the temporal context transformer effectively extracts the context information of different frame images. The aggregate feature will reshape to its original dimension before being sent into the decoder.

3. Experimental Results

In this section, we compare the performance of the proposed ensemble Temporal Context for Endoscopy Artefact Segmentation and Detection Farmworke and state-of-the-art model were compared in the segmentation and detection of endoscopy artefact.

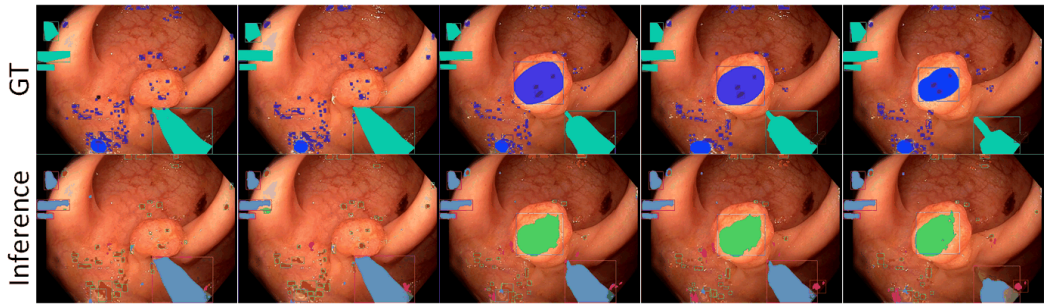

Figure 3: Example of sequential endoscopy artefact image segmentation and detection results.

Table 1
Temporal context transformer layer number comparative experiment.

Model	N	$Dice$	$Jaccard$	PA
UNet	0	0.525	0.402	0.872
	1	0.635	0.491	0.892
	2	**0.653**	**0.513**	**0.897**
	3	0.607	0.469	0.895
Model	N	mAP_{mean}	mAP_{50}	mAP_{75}
Faster R-CNN	0	0.232	0.464	0.208
	1	0.305	0.554	0.309
	2	**0.317**	**0.563**	**0.321**
	3	0.288	0.523	0.272

Data details and preparation. Our model mainly used the EndoCV2022 challenge dataset [17] for endoscopic images for Endoscopy Artefact Detection in this work. Endoscopic surgical instruments include five categories: nonmucosa, artefact, saturation, specularity, bubbles. EndoCV launched this as an extension to the previous artefact detection and segmentation challenges [21, 22] with dataset specific to the colonoscopy. The dataset contains 24 endoscopic videos sequence for EAD sub-challenge with total 1,449 endoscopic images. We split the dataset into 80% sequence for training and 20% sequence for validation. For the segmentation task, we used Dice coefficient, Jaccard coefficient and PA for evaluation. For the detection task, we used mAP with different thresholds for evaluation.

Implementation details. The deep models are implemented based on PyTorch and trained on an NVIDIA Tesla V100 GPU. surgical instrument segmentation model using SGD optimizer with a learning rate of 10^{-4}. surgical instrument detection model base on mmdetetcion and using SGD optimizer with a learning rate of 10^{-2}. The batch size is set to 2 and use a sliding window of length L to sample subsequences in the original sequence, while input sequence images are resized to 960×540. Since the input are image sequences, the batch size was relatively small. In addition, we used conventional inversion, affine transformation, contrast and other methods to enhance the data of the training set. In order to demonstrate the effectiveness of the method, we do not use TTA or multi-model fusion and other post-processing means, but only use a single model for test set prediction.

Table 2
Structural ablation of temporal context transformer

Model	$TCT_{w/o}$	$Dice$	$Jaccard$	PA
UNet		0.525	0.402	0.872
	√	**0.635**	**0.491**	**0.892**
DANet		0.651	0.597	0.923
	√	**0.773**	**0.660**	**0.944**
PraNet		0.716	0.676	0.936
	√	**0.815**	**0.721**	**0.961**
Model	$TCT_{w/o}$	mAP_{mean}	mAP_{50}	mAP_{75}
Faster R-CNN		0.232	0.464	0.208
	√	**0.317**	**0.563**	**0.321**
Cascade RCNN		0.336	0.579	0.347
	√	**0.395**	**0.611**	**0.401**
Swin Transformer		0.356	0.598	0.364
	√	**0.403**	**0.613**	**0.421**

We first compared the influence of the number N of TCT on the model performance through comparative experiments. The results are shown in Table 1. From the experimental results, it can be seen that the model has the best effect except when N is 2, and the model will overfit when N is too large. To verify the effectiveness of our method, we perform a comprehensive comparison with state-of-the-art segmentation and detection methods, segmentation methods including UNet, DANet, PraNet, detection methods including Faster RCNN, Cascade RCNN, Swin Transformer, as shown in Table 2. Specifically, The performance of each SOTA model has been steadily improved after being converted to our method. We visualized an example of an inferential endoscope sequence image of a set of models, as shown in Fig.3. the Dice, Jaccar, PA of segmentation task model has been improved by 9%-12%, 5%-9% and 2%-3%. For detection tasks, the model's mAP improved by 5%-8%. And it is effective for different types of methods, which proves that our method is robust and applicable.

References

[1] Q. L. Zhe Guo, Ruiyao Zhang, et al., Global cancer statistics, 2012„ Ca A Cancer Journal for Clinicians 65 (2013) 87–108.

[2] P. L. Reiko Nishihara, Kana Wu, et al., Long-term colorectal-cancer incidence and mortality after lower endoscopy (2018).

[3] Y. Mori, S. Kudo, Detecting colorectal polyps via machine learning, Nature Biomedical Engineering 2 (2018) 713–714.

[4] K. T. Teppei Kanayama, Yusuke Kurose, et al., Gastric Cancer Detection from Endoscopic Images Using Synthesis by GAN, MICCAI 2019, Part V, 2019.

[5] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, G. Hamarneh, Deep semantic segmentation of natural and medical images: a review, Artif. Intell. Rev. 54 (2021) 137–178. URL: https://doi.org/10.1007/s10462-020-09854-1. doi:10.1007/s10462-020-09854-1.

[6] G. Zamanakos, L. T. Tsochatzidis, A. Amanatiadis, I. Pratikakis, A comprehensive survey of lidar-based 3d object detection methods with deep learning for autonomous driving, Comput. Graph. 99 (2021) 153–181. URL: https://doi.org/10.1016/j.cag.2021.07.003. doi:10.1016/j.cag.2021.07.003.

[7] O. Ronneberger, Invited talk: U-net convolutional networks for biomedical image segmentation, in: K. H. Maier-Hein, T. M. Deserno, H. Handels, T. Tolxdorff (Eds.), Bildverarbeitung für die Medizin 2017 - Algorithmen - Systeme - Anwendungen. Proceedings des Workshops vom 12. bis 14. März 2017 in Heidelberg, Informatik Aktuell, Springer, 2017, p. 3. URL: https://doi.org/10.1007/978-3-662-54345-0_3. doi:10.1007/978-3-662-54345-0_3.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Medical Imaging 39 (2020) 1856–1867. URL: https://doi.org/10.1109/TMI.2019.2959609. doi:10.1109/TMI.2019.2959609.

[9] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, K. H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2020) 203â211. URL: http://dx.doi.org/10.1038/s41592-020-01008-z. doi:10.1038/s41592-020-01008-z.

[10] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3146–3154. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.html. doi:10.1109/CVPR.2019.00326.

[11] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII, volume 11211 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 833–851. URL: https://doi.org/10.1007/978-3-030-01234-2_49. doi:10.1007/978-3-030-01234-2_49.

[12] D. Fan, G. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Pranet: Parallel reverse attention network for polyp segmentation, in: A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI, volume 12266 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 263–273. URL: https://doi.org/10.1007/978-3-030-59725-2_26. doi:10.1007/978-3-030-59725-2_26.

[13] S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. URL: https://doi.org/10.1109/TPAMI.2016.2577031. doi:10.1109/TPAMI.2016.2577031.

[14] Z. Cai, N. Vasconcelos, Cascade R-CNN: delving into high quality object detection, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, Computer Vision Foundation / IEEE Computer Society, 2018, pp. 6154–6162. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Cai_Cascade_R-CNN_Delving_CVPR_2018_paper.html. doi:10.1109/CVPR.2018.00644.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows (2021).

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), NIPS 2017, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[17] S. Ali, D. Jha, N. Ghatwary, S. Realdon, R. Cannizzaro, O. E. Salem, D. Lamarque, C. Daul, K. V. Anonsen, M. A. Riegler, et al., Polypgen: A multi-

center polyp detection and segmentation dataset for generalisability assessment, arXiv preprint arXiv:2106.04463 (2021).

[18] S. Ali, N. Ghatwary, D. Jha, E. Isik-Polat, G. Polat, C. Yang, W. Li, A. Galdran, M.-Á. G. Ballester, V. Thambawita, et al., Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge, arXiv preprint arXiv:2202.12031 (2022).

[19] Q. Wang, L. Zhou, Y. Yao, Y. Wang, J. Li, W. Yang, An interconnected feature pyramid networks for object detection, J. Vis. Commun. Image Represent. 79 (2021) 103260. URL: https://doi.org/10.1016/j.jvcir.2021.103260. doi:10.1016/j.jvcir.2021.103260.

[20] N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-nms - improving object detection with one line of code, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 5562–5570. URL: https://doi.org/10.1109/ICCV.2017.593. doi:10.1109/ICCV.2017.593.

[21] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, S. Albarqouni, X. Wang, C. Wang, S. Watanabe, I. Oksuz, Q. Ning, S. Yang, M. A. Khan, X. W. Gao, S. Realdon, M. Loshchenov, J. A. Schnabel, J. E. East, G. Wagnieres, V. B. Loschenov, E. Grisan, C. Daul, W. Blondel, J. Rittscher, An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy, Scientific Reports 10 (2020). URL: https://doi.org/10.1038%2Fs41598-020-59413-5. doi:10.1038/s41598-020-59413-5.

[22] S. Ali, M. Dmitrieva, N. Ghatwary, S. Bano, G. Polat, A. Temizel, A. Krenzer, A. Hekalo, Y. B. Guo, B. Matuszewski, M. Gridach, I. Voiculescu, V. Yoganand, A. Chavan, A. Raj, N. T. Nguyen, D. Q. Tran, L. D. Huynh, N. Boutry, S. Rezvy, H. Chen, Y. H. Choi, A. Subramanian, V. Balasubramanian, X. W. Gao, H. Hu, Y. Liao, D. Stoyanov, C. Daul, S. Realdon, R. Cannizzaro, D. Lamarque, T. Tran-Nguyen, A. Bailey, B. Braden, J. E. East, J. Rittscher, Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy, Medical Image Analysis 70 (2021) 102002. URL: https://doi.org/10.1016%2Fj.media.2021.102002. doi:10.1016/j.media.2021.102002.