Colonoscopy Landmark Detection Using Vision Transformers

Aniruddha Tamhane(✉), Tse'ela Mida, Erez Posner, and Moshe Bouhnik

Intuitive Surgical, Inc., 1020 Kifer Road, Sunnyvale, CA, USA
{aniruddha.tamhane,tseela.mida,erez.posner,moshe.bouhnik}@intusurg.com

Abstract. Colonoscopy is a routine outpatient procedure used to examine the colon and rectum for any abnormalities including polyps, diverticula and narrowing of colon structures. A significant amount of the clinician's time is spent in post-processing snapshots taken during the colonoscopy procedure, for maintaining medical records or further investigation. Automating this step can save time and improve the efficiency of the process. In our work, we have collected a dataset of 120 colonoscopy videos and 2416 snapshots taken during the procedure, that have been annotated by experts. Further, we have developed a novel, vision-transformer based landmark detection algorithm that identifies key anatomical landmarks (the appendiceal orifice, ileocecal valve/cecum landmark and rectum retroflexion) from snapshots taken during colonoscopy. Our algorithm uses an adaptive gamma correction during preprocessing to maintain a consistent brightness for all images.We then use a vision transformer as the feature extraction backbone and a fully connected network based classifier head to categorize a given frame into four classes: the three landmarks or a non-landmark frame. We compare the vision transformer (ViT-B/16) backbone with ResNet-101 and ConvNext-B backbones that have been trained similarly. We report an accuracy of 82% with the vision transformer backbone on a test dataset of snapshots.

Keywords: Colonoscopy · Vision transformer · Landmark detection

1 Introduction

Colorectal cancer (CRC) is among the leading causes of death worldwide [4]. In the United States alone, 161,470 individuals are estimated to be diagnosed with CRC and 54,250 individuals are estimated to die from CRC in 2022 [26]. Colorectal cancer incidence rates have been increasing among screening-age individuals aged 65 years and older by 1% per year [27]. Early onset CRC rates have also been on the rise among the patients under the recommended screening age (50 years). Early screening for colorectal abnormalities is associated with a 67% reduction in mortality from CRC [9]. Colonoscopy being the gold standard for CRC screening [13] plays a critical role in mitigating risk.

L. Manfredi et al. (Eds.): ISGIE 2022/GRAIL 2022, LNCS 13754, pp. 24–34, 2022.
https://doi.org/10.1007/978-3-031-21083-9_3

Snapshots taken during the colonoscopy are a critical yet time-consuming part of the post-procedural diagnosis and documentation. Physicians typically take snapshots of key colon landmarks such as the Appendiceal Orifice (AO), Ileocecal Valve (ICV), Cecum landmark (Cec) and certain findings such as polyps, diverticula, or routine procedural steps such as a Rectum Retroflexion (RecRF), as recommended by the American Gastroenterological Institute [7]. The snapshots are useful in the post-procedural phase to serve as a medical record of the highlights of the colonoscopy and the patient's colonic health or for assessing the extent of the procedure by capturing a snapshot of the appendiceal orifice and ileocecal valve [21].

It has been reported in [19] that a significant amount of a clinician's time is spent maintaining Electronic Health Records. With the increase in demand for colonoscopy procedures, there is a need for improving the efficiency to save the colonoscopy clinician's time. There have been multiple robust, highly accurate and efficient approaches developed for polyp detection [18,23,24]. However, there has been a limited amount of research on landmark detection. To the best of our knowledge, the algorithms developed by [2,16] have been the only attempts at detecting the appendiceal orifice (using classical and deep learning techniques respectively). The deep-learning technique developed by [14] to detect the hepatic and splenic flexure, is the only multi-landmark detection algorithm for colons. We believe that this scarcity of available literature may be due to a lack of availability of expert annotated datasets of colon landmarks and the inherent difficulty of the task due to: *1)* intra-colon (patient) similarity between different regions, *2)* inter-colon (patient) variability in the anatomical structures of the same region of the colon and *3)* non-ideal photometric conditions of the snapshots (due to poor focus, blur, reflections on the colon walls, occlusions by fluids, polyps etc.) Thus, there is a need for developing a robust technique that can accurately identify anatomical landmarks in the colon across multiple patients, that has been rigorously tested on a dataset containing colonoscopy snapshots that are representative of the typical clinical setting. Further, it is important to design a data-efficient training framework that can demonstrably generalize across different anatomies.

We propose a vision transformer based training framework that enables a model trained on videos (which are cheaper to annotate) to be adapted for snapshots. In our work, we address the following problems pertaining our task: *1)* adaptation to differences in data distribution from video-annotations to snapshots *2)* extreme class imbalance, *3)* poor photometric conditions and *4)* inconsistent annotations from experts.

2 Related Work

A large body of work on the application of statistical, physics-based analysis and machine-learning techniques on colonoscopy has accumulated over the years primarily focusing on the detection of polyps and to a lesser extent, colon landmarks. We review the following categories of scientific literature relevant to our work:

2.1 Landmark Detection

Fast, reliable techniques of detecting anatomical landmarks are crucial to medical image analysis. Landmark detection in ultrasound and CT scans is a well explored field, with research on detecting landmarks to utilizing them for organ segmentation [6,11,30,31]. Detecting landmarks in endoscopy and colonoscopy has a smaller yet broader research focusing on identifying different landmarks and regions as a part of the endo-/colonoscopy process. In [2], a shape-based feature extraction model combined with K-Means clustering was used to detect the appendiceal orifice in colonoscopy videos. Since this method relies on edge-based shape detection, there is a possibility of it not working on blurry images, which are characteristic of typical colonoscopy snapshots. A deep-learning based approach was proposed in [1] for detecting the anatomical regions (e.g. stomach, oesophagus etc.) from capsule endoscope frames. This demonstrated the efficacy of deep networks to correctly identify anatomical regions from a single endoscopy frame. The first major attempt at identifying certain colon landmarks from colonoscopy frames using deep neural networks was made by [3]. They trained a large 2D CNN based neural network to classify a given frame as either one of splenic flexure, hepatic flexure or sigmoidal colon junction. Their approach relies on removing blurred frames using a heuristic, and on testing the model on non-overlapping frames from the videos common to the training set.

2.2 Visual Feature Backbones and Optimizers

Convolutional Neural Network based architectures such as the VGG-16 [28] and ResNet101 [12] have traditionally been the most effective and widely used visual feature extraction architectures. The ConvNext [17] is the latest state-of-the-art CNN-based architecture. On the other hand, the transformer architecture [29], which is the standard architecture in Natural Language Processing, has now been adapted for vision-related tasks in [8] showing promising results. Due to the fundamentally different mechanisms of transformer-based (attention) and CNN-based architectures (learned filters), we decide to compare both types of architectures for our task. For our primary model, we use a Vision Transformer pre-trained on the ImageNet dataset as the visual feature extraction backbone. We also independently train a ResNet-101 and a ConvNext based model for comparison. The choice of optimizer used directly affects the optimization landscape impacting the accuracy and ability to generalize, as show in [5]. We use a Sharpness Aware Minimization (SAM) [10] approach to optimizing neural networks due to its positive impact on the accuracy as well as producing semantically meaningful attention maps in case of transformers.

3 Data Collection

We have collected and annotated 120 colonoscopy videos and 2416 snapshots that have been used for training and evaluating our algorithm respectively. We describe the annotation process, training dataset and snapshots dataset in the following subsections.

3.1 Annotations and Cross-Validation

We have annotated the videos on a frame-level and have cross-validated the annotations between the medical experts. This ensures a clinically accurate dataset that has fine-grain annotations with fewer human errors. We have followed the same procedure while annotating the training videos as well as the snapshots dataset. Our annotation methodology is as follows: we separate videos for the training data (which will be further split into validation and testing sets) and the snapshots dataset. Separating the data on a video-level is critical to ensure that the model generalizes well to all the anatomical variations found in colons. Each of the videos in the training datasets is then labelled on a frame-level by two medical students independently. Only the frames with a consensus between the two annotators are chosen for training and the rest are discarded. On the other hand, each of the videos in the snapshots dataset was examined by a senior medical expert to extract snapshots, as they would in a clinical setting. Each of these snapshots was then labelled independently by two senior medical experts, and a similar consensus-based cross-validation heuristic was used to select the snapshots with matching annotations from the two experts.

Table 1. Snapshots and test dataset label distribution

Label	Number of frames (Snapshot)	Number of frames (test)
Appendiceal Orifice	518	776
Ileocecal Valve/Cecum Landmark	132	133
Rectum retroflexion	716	140
Other	1050	1488

3.2 Snapshots Dataset

Our snapshots dataset contains 2416 snapshots collected from over 500 videos (separate from the training pool of 120 videos), identified and annotated by clinicians as described in Subsect. 3.1. A snapshot is a video frame that contains the anatomical/procedural feature of interest in reasonable focus, as identified by a medical specialist in a clinical setting. Each of the snapshots have been annotated according to the following labels: Appendiceal Orifice (AO), Ileocecal Valve (ICV)/Cecum Landmark (Cec), Rectum Retroflexion (RecRF) and Other, which are shown with examples in Fig. 1. Since the Ileocecal Valve and the Cecum Landmark typically co-occur in snapshots due to their anatomical proximity, we combine them into a single label. Both of the first two labels describes the corresponding anatomical landmark. RecRF refers to the procedural action of retroflexion in the rectum i.e. bending the colonoscope backwards to inspect the rectum. Any other anatomical finding such as polyps, inflammation or general anatomical markers have been labelled as “Other”. A breakdown of the number of frames per class has been given in Table 1.

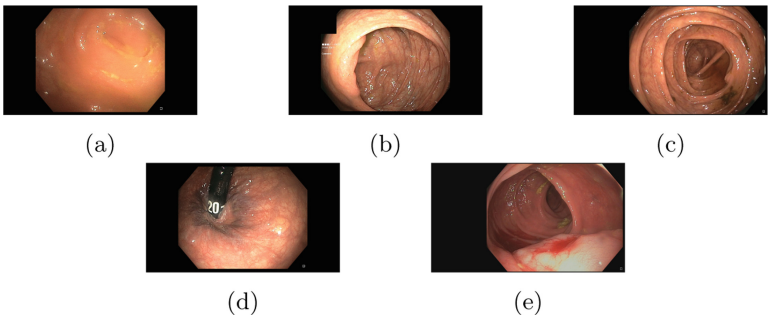


Fig. 1. Sample snapshots with following annotations: Appendiceal Orifice (1a), Ileocecal Valve (1b), Cecum Landmark (1c), Rectum Retroflexion (1d), Other (1e)

3.3 Training Dataset

Our training dataset has 120 videos constituting of 2,000,000 frames in all, that were annotated and cross-validated as described in Subsect. 3.1. We face an extreme label imbalance, with a majority of frames (>95%) belonging to a non-landmark (Other) class, and the minority containing a landmark of interest. We balance the dataset as part of our training and evaluation (to get a distribution similar to the snapshots dataset) as described in Sect. 6.

4 Problem Definition

We define our problem as follows: identify a function $f : C \times H \times W \rightarrow J$ to classify an image frame F as one of the landmark classes $j \in \{\text{AO}, \text{ICV/Cec}, \text{RecRF}, \text{Other}\}$ such that $f(F_{ij}) = j, \quad \forall i \in \mathcal{S}, j \in J$. Here, $\mathcal{S}, J$ denote the set of snapshots and class labels respectively. We approximate f using a deep neural network due to their proven capacity for modeling image data. We thus reduce our problem to finding the optimal weights θ^* for the following empirical loss ($\mathcal{L}$) :

$$\theta^* = \arg\min_{\theta} \sum_{i,j} \mathcal{L}(f(h(F_{ij})|\theta), j) \tag{1}$$

Here, h refers to a general data preprocessing function. Our framework supports any loss function $\mathcal{L}$ that is a distance metric between the predicted probability distribution and the true labels. Based on our experiments, we choose a Kullback-Leibler Divergence [15] as the loss function $\mathcal{L}$.

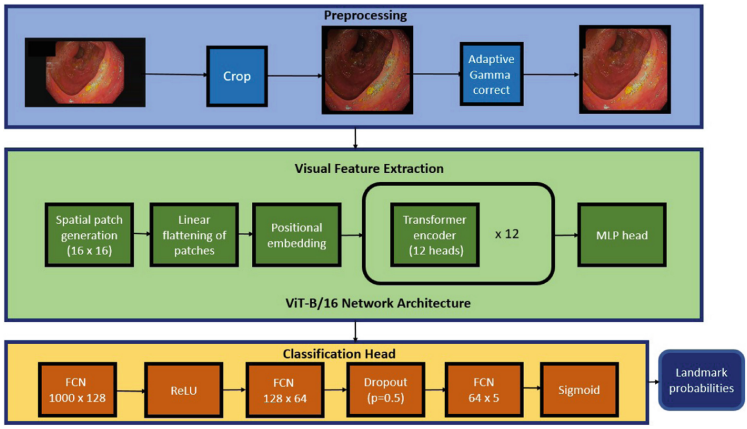


Fig. 2. Landmark detection pipeline architecture

5 Architecture

Our algorithm consists of three primary parts: *1)* image preprocessing, *2)* visual feature extraction and *3)* classification head. The image preprocessing consists of an auto-cropping step to remove dark edges that are an artifact of the colonoscopy software itself, and auto-correct the brightness using gamma correction. Since the brightness varies considerably during a colonoscopy, we use an adaptive gamma correction algorithm described in [25]. We use a pretrained Vision Transformer (ViT-B/16) as the visual feature extraction backbone in our primary model. We also experiment with other CNN based architectures (ResNet101 and ConvNext-B) that were identically pretrained on the ImageNet dataset and benchmark their performances. Finally, we a use a Fully Connected Network (FCN) based classifier head to compute the label probabilities from the feature vector generated by the backbone. A high-level overview of the architecture is given in Fig. 2.

6 Training Pipeline

We design our framework to train a model on annotated videos so that it performs well on clinically selected snapshots. Snapshots are different from video frames because they are hand-picked by clinicians in the following regards: they have a different distribution of landmarks and have a different photometric quality. We address this gap in the training and evaluation data using:

1. *Cross-validation*: Cross-validating frames as explained in Sect. 3.1 reduces the possibility of annotation error and inclusion of poor quality frames in the training. This bridges the gap in data quality between snapshots and videos.
2. *Domain-specific sampling*: We artificially construct a training set that has a label distribution similar to the snapshots dataset by randomly sub-sampling the frames using a Bernoulli process, described in Eqs. 2, 3. Thus, a frame F_{ij} is included in the training set if $Z_{ij} = 1$. Here, $\mathcal{S}$, $\mathcal{T}$ are the snapshots and training sets respectively. $|\Gamma|$ denotes the cardinality of any set Γ.

$$Z_{ij} \sim \text{Bernoulli}(p_j) \tag{2}$$

$$p_j = min\left(\frac{|\bigcup_{i\in\mathcal{S},k=j} F_{ik}|}{|\bigcup_{i\in\mathcal{S},k} F_{ik}|} \Big/ \frac{|\bigcup_{i\in\mathcal{T},k=j} F_{ik}|}{|\bigcup_{i\in\mathcal{T},k} F_{ik}|}, 1\right) \tag{3}$$

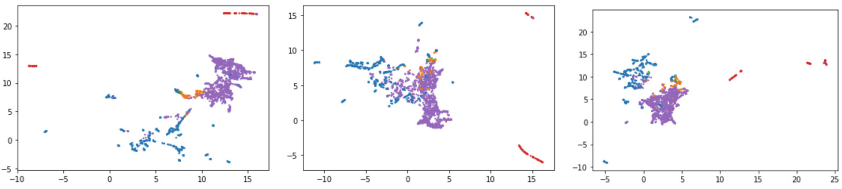
   We repeat the sampling (with replacement) at the beginning of every epoch to maximally cover the downsampled frames.
3. *Sharpness-Aware Minimization Optimizer*: Learning anatomically relevant features and ignoring features generated by varying photometric conditions, specific clinical conditions etc. is critical to generalizability across multiple patient anatomies. We observe that using a SAM optimization scheme as described in [10] for training the neural networks helps learn such a robust model.

7 Results

We have trained Vision Transformer (ViT-B/16), ResNet-101 and ConvNext-B based models using our framework and evaluated the results on our snapshots dataset. We tabulate the corresponding accuracy and the class-wise precision, recall scores in Table 2. We also plot 2D U-MAP [20] embeddings of the vision backbone representations for images from our balanced test dataset in Fig. 3. We report the test dataset statistics in Table 1. We see that the vision transformer based model outperforms the other two on most metrics reported in Table 2. This is also corroborated by the comparatively well-separated clusters in Fig. 3. We believe that the inherent shape bias of vision transformers, as reported in [22], makes it more suitable than CNN-based architectures for landmark detection, since landmarks are reliably identified by their shape regardless of texture.

Table 2. Recall, precision scores and overall accuracy on snapshots dataset

Class	Metric	ViT-B/16 (Main)	ResNet-101	ConvNext-B
Overall	Accuracy	**81.84%**	73.06%	60.45%
AO	Recall	68.15%	69.69%	**75.09%**
	Precision	**76.41%**	55.36%	57.12%
ICV/Cec	Recall	**89.43%**	75.33%	88.11%
	Precision	51.26%	**55.52%**	24.84%
RecRF	Recall	**96.09%**	86.31%	88.12%
	Precision	**98.29%**	97.48%	95.03%
Other	Recall	**77.24%**	65.05%	28.39%
	Precision	**85.10%**	74.48%	82.55%



(a) ViT-B/16 (b) ConvNext-B (c) ResNet-101

Fig. 3. U-MAP embeddings of vision backbone representations with the color scheme: AO (Blue), ICV (Orange), Cec (Green), RecRF (Red), Purple (Other) (Color figure online)

8 Inference and Future Work

We achieve an overall landmark classification accuracy of 81.84% on a snapshot dataset of clinically relevant colon landmarks using a vision transformer backbone. We observe that a transformer based backbone outperforms other state-of-the-art CNN-based backbones such as ResNet-101 and ConvNext-B. We can visually see that well-separated representations on an independent, balanced test set imply a higher accuracy in Fig. 3. This may be due to the transformer's inherently higher shape bias as reported by [22]. We hypothesize thus, since the accuracy trend is not completely explained by the number of parameters, with ViT-B/16 (86.6M) and ConvNext-B (89M) having a comparable number of parameters.

Further, the Rectum Retroflexion class has the highest precision and recall scores as well as the best separation on the U-MAP plots. This is because most RecRF frames are characterized by the presence of a metallic/plastic tube indicating the inversion of the colonoscope head. We further observe that the precision for AO and ICV classes is relatively lower. This is also evidenced by the

poorer separation of the corresponding clusters in Fig. 3. This can be explained by the visual similarity between these two landmarks and other parts of the colon (labelled "Other"), making it a challenging task. Thus, we can conclude from our results that detecting subtle anatomical features (such as a cecum landmark) as opposed to specific shapes (such as the retroflexion tube) is challenging for the vision backbone.

Finally, we believe incorporating temporal information in our future work will help remove false positives for both these classes and improve precision. We also believe that more complex training techniques such as active learning, self-supervised pre-training can further improve the quality of features learned by the vision backbone and improve accuracy. So, we plan on incorporating them in our future pipeline. We also plan on including more landmark classes such as polyps and diverticula in the future.

References

1. Adewole, So., et al.: Deep learning methods for anatomical landmark detection in video capsule endoscopy images. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) FTC 2020. AISC, vol. 1288, pp. 426–434. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-63128-4_32
2. Cao, Y., Liu, D., Tavanapong, W., Wong, J., Oh, J., De Groen, P.C.: Automatic classification of images with appendiceal orifice in colonoscopy videos. In: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2349–2352. IEEE (2006)
3. Che, K., et al.: Deep learning-based biological anatomical landmark detection in colonoscopy videos. arXiv preprint arXiv:2108.02948 (2021)
4. Chen, J., et al.: Cause of death among patients with colorectal cancer: a population-based study in the united states. Aging (Albany NY) **12**(22), 22927 (2020)
5. Chen, X., Hsieh, C.J., Gong, B.: When vision transformers outperform resnets without pretraining or strong data augmentations. arXiv preprint arXiv:2106.01548 (2021)
6. Chowdhury, A.S., Yao, J., VanUitert, R., Linguraru, M.G., Summers, R.M.: Detection of anatomical landmarks in human colon from computed tomographic colonography images. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)
7. Cooper, J.A., Ryan, R., Parsons, N., Stinton, C., Marshall, T., Taylor-Phillips, S.: The use of electronic healthcare records for colorectal cancer screening referral decisions and risk prediction model development. BMC Gastroenterol. **20**(1), 1–16 (2020)
8. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Doubeni, C.A., et al.: Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: a large community-based study. Gut **67**(2), 291–298 (2018). https://doi.org/10.1136/gutjnl-2016-312712, https://gut.bmj.com/content/67/2/291
10. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412 (2020)

11. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 229–237. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46726-9_27
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
13. Issa, I.A., Noureddine, M.: Colorectal cancer screening: an updated review of the available options. World J. Gastroenterol. **23**(28), 5086 (2017)
14. Jheng, Y.C., et al.: A novel machine learning-based algorithm to identify and classify lesions and anatomical landmarks in colonoscopy images. Surg. Endoscopy 1–11 (2021)
15. Kim, T., Oh, J., Kim, N., Cho, S., Yun, S.Y.: Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation. arXiv preprint arXiv:2105.08919 (2021)
16. Lebedev, A., Khryashchev, V., Kazina, E., Zhuravleva, A., Kashin, S., Zavyalov, D.: Automatic identification of appendiceal orifice on colonoscopy images using deep neural network. In: 2020 IEEE East-West Design & Test Symposium (EWDTS), pp. 1–5. IEEE (2020)
17. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. arXiv preprint arXiv:2201.03545 (2022)
18. Mamonov, A.V., Figueiredo, I.N., Figueiredo, P.N., Tsai, Y.H.R.: Automated polyp detection in colon capsule endoscopy. IEEE Trans. Med. Imaging **33**(7), 1488–1502 (2014)
19. McDonald, C.J., Callaghan, F.M., Weissman, A., Goodwin, R.M., Mundkur, M., Kuhn, T.: Use of internist's free time by ambulatory care electronic medical record systems. JAMA Internal Med. **174**(11), 1860–1863 (2014)
20. McInnes, L., Healy, J., Melville, J.: Umap: uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
21. Morelli, M.S., Miller, J.S., Imperiale, T.F.: Colonoscopy performance in a large private practice: a comparison to quality benchmarks. J. Clin. Gastroenterol. **44**(2), 152–153 (2010)
22. Morrison, K., Gilby, B., Lipchak, C., Mattioli, A., Kovashka, A.: Exploring corruption robustness: inductive biases in vision transformers and mlp-mixers. arXiv preprint arXiv:2106.13122 (2021)
23. Park, S.Y., Sargent, D., Spofford, I., Vosburgh, K.G., Yousif, A., et al.: A colon video analysis framework for polyp detection. IEEE Trans. Biomed. Eng. **59**(5), 1408–1418 (2012)
24. Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I.: Toward real-time polyp detection using fully CNNS for 2d gaussian shapes prediction. Med. Image Anal. **68**, 101897 (2021)
25. Rahman, S., Rahman, M.M., Abdullah-Al-Wadud, M., Al-Quaderi, G.D., Shoyaib, M.: An adaptive gamma correction for image enhancement. EURASIP J. Image Video Process. **2016**(1), 1–13 (2016). https://doi.org/10.1186/s13640-016-0138-1
26. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA: A Cancer J. Clin. **72**, 7–33 (2022)
27. Siegel, R.L., et al.: Colorectal cancer statistics, 2020. CA: A Cancer J. Clin. **70**(3), 145–164 (2020)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

29. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
30. Zhou, S.K., et al.: A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. In: Proceedings of the IEEE (2021)
31. Zhou, S.K., Xu, Z.: Landmark detection and multiorgan segmentation: representations and supervised approaches. In: Handbook of Medical Image Computing and Computer Assisted Intervention, pp. 205–229. Elsevier (2020)