Loose Lesion Location Self-supervision Enhanced Colorectal Cancer Diagnosis

Tianhong Gao[1], Jie Song[1], Xiaotian Yu[1], Shengxuming Zhang[1], Wenjie Liang[2], Hongbin Zhang[3], Ziqian Li[1], Wenzhuo Zhang[2], Xiuming Zhang[1], Zipeng Zhong[1], Mingli Song[1], and Zunlei Feng[1(✉)]

[1] Zhejiang University, Hangzhou, China
zunleifeng@zju.edu.cn
[2] The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China
[3] Yiwu Central Hospital, Yiwu, China

Abstract. Early diagnosis of colorectal cancer (CRC) is crucial for improving survival and quality of life. While computed tomography (CT) is a key diagnostic tool, manually screening colon tumors is time-consuming and repetitive for radiologists. Recently, deep learning has shown promise in medical image analysis, but its clinical application is limited by the model's unexplainability and the need for a large number of finely annotated samples. In this paper, we propose a loose lesion location self-supervision enhanced CRC diagnosis framework to reduce the requirement of fine sample annotations and improve the reliability of prediction results. For both non-contrast and contrast CT, despite potential deviations in imaging positions, the lesion location should be nearly consistent in images of both modalities at the same sequence position. In addition, lesion location in two successive slices is relatively close for the same modality. Therefore, a self-supervision mechanism is devised to enforce lesion location consistency at both temporal and modality levels of CT, reducing the need for fine annotations and enhancing the interpretability of diagnostics. Furthermore, this paper introduces a mask correction loopback strategy to reinforce the interdependence between category label and lesion location, ensuring the reliability of diagnosis. To verify our method's effectiveness, we collect data from 3,178 CRC patients and 887 healthy controls. Experiment results show that the proposed method not only provides reliable lesion localization but also enhances the classification performance by 1-2%, offering an effective diagnostic tool for CRC. Code is available at https://github.com/Gaotianhong/LooseLocationSS.

Keywords: Colorectal cancer · Computed tomography · Loose location self-supervision · Mask correction loopback · Reliability.

1 Introduction

Colorectal cancer (CRC), accounting for about 10% of all cancer cases, is the third most common cancer and the second leading cause of cancer-related deaths

globally [22]. In modern medical diagnostics, computer tomography (CT) has become an important means for CRC screening due to its high resolution and non-invasive characteristics [2,21]. However, traditional CT image analysis relies on manual screening by radiologists, which is time-consuming, labor-intensive and subjective. Therefore, automated medical image analysis methods such as deep learning technology have become a research hotspot, aiming to improve the efficiency and accuracy of diagnosis [3,17,15,30].

Inspired by the successful application of deep learning, several works adopted neural network to diagnose CRC in CT images [25,8,19,28,1,29,6], primarily focusing on classification, detection and segmentation. The classification methods [25,8] trained 3D CNN to determine the presence of CRC in CT scans.In lesion detection, Sahoo *et al.* [19] used RetinaNet [13] and YOLO [24] to localize lesion in CT images. Yao *et al.* [28] developed a deep learning model for detection of CRC and compared it with radiologists. In lesion segmentation, Akilandeswari *et al.* [1] adopted ResNet-enabled CNN [7] to achieve complete boundary segmentation of the colon cancer region. Yao *et al.* [29] proposed a topology-aware approach for automated colorectum and CRC segmentation in routine abdominal CT scans. Han *et al.* [6] customized 3D U-Net of nnU-Net [10] to simultaneously detect and segment lesion.

Although the CRC classification method does not require fine labeling, the black-box nature of deep learning models lead to insufficient explanation. While detection and segmentation methods provide a degree of interpretability, the need for extensive labeling restricts the clinical application of deep learning.

In this paper, we propose a loose lesion location self-supervision and devise a lesion location mask correction loopback mechanism, which not only reduces the need for fine annotations, but also provides a diagnostic basis and improves the interpretability of model. The overall framework is shown in Fig. 1. Considering the spatial continuity of lesion location across successive slices in the same modality, we adopt the temporal consistency constraint. For non-contrast and contrast CT, despite potential deviations in imaging positions, images from both modalities at the same sequence position should display roughly same lesion area, leading to the adoption of modality consistency constraint. Furthermore, through lesion location loopback, the patch identified as lesion is mapped back to the original CT image and masked, and then sent to the classification branch to establish a strong dependence between classification results and lesion location, ensuring the reliability of CRC diagnosis.

A colorectal CT scanning dataset was collected from 3,178 CRC patients and 887 healthy controls. Each sample contains both non-contrast and contrast CT with continuous image sequences in each modality. Experiment results show that our method requires only 8.3% fine annotations to achieve localization performance comparable to fully supervised methods, enhances diagnostic accuracy and provides reliable lesion location, thus offering an effective diagnostic tool.

Our main contribution can be summarized as follows: 1) We propose a loose lesion location self-supervision mechanism by constraining the temporal and cross-modal misalignment of lesion locations, which achieves accurate le-

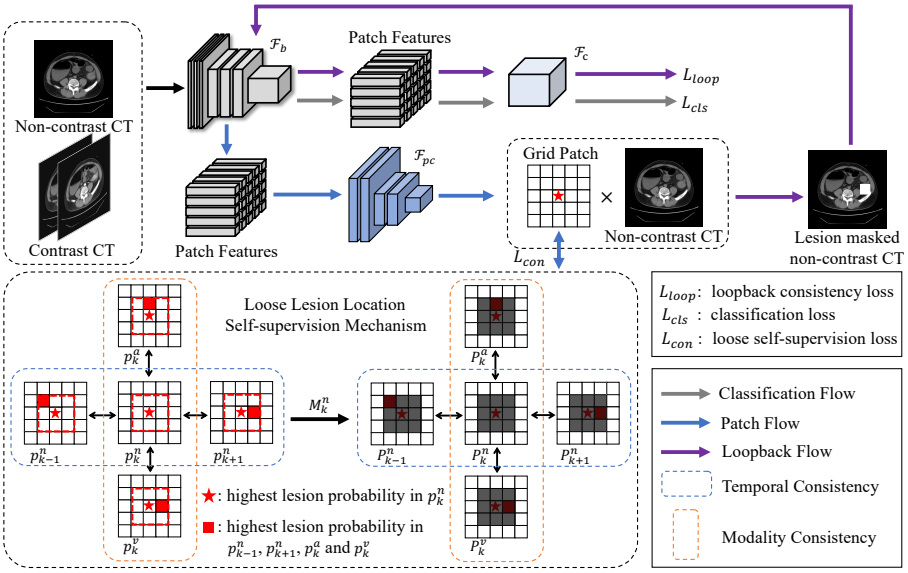

Fig. 1. The framework of loose lesion location self-supervision enhanced CRC diagnosis, which is composed of image classification branch $\mathcal{F}_b \circ \mathcal{F}_c$ and patch classification branch $\mathcal{F}_b \circ \mathcal{F}_{pc}$. The image classification branch enables the model to distinguish between CRC and healthy samples. The patch classification branch achieves accurate lesion localization through loose location self-supervision constraint, which reduces the need for labeled samples and enhances interpretability. Additionally, the mask correction loopback mechanism constrains the strong dependence of classification results and lesion location.

sion localization with minimal finely labeled data. 2) The lesion location mask correction loopback mechanism is devised to enhance the consistency between category label and lesion location, which improves the reliability of diagnostic results. 3) The proposed method achieves excellent classification and localization performance with limited annotations on the collected dataset.

2 Method

Given the limitations of fully supervised diagnostic methods, which require extensive annotations and suffer from poor interpretability, we propose a loose lesion location self-supervision enhanced CRC diagnosis framework, illustrated in Figure 1. By analyzing the slight misalignment consistency of lesion location, the temporal and modality self-supervision strategy is proposed. Combined with fuzzy classification loss function, the lesion sites are accurately located. Furthermore, mask correction loopback mechanism, which strengthens the consistency between the predicted results and the lesion location by masking the lesion area, increases the reliability.

The framework is composed of an image classification branch $\mathcal{F}_b \circ \mathcal{F}_c$ and a patch classification branch $\mathcal{F}_b \circ \mathcal{F}_{pc}$. The two branches share the same backbone $\mathcal{F}_b$. The $\mathcal{F}_c$ and $\mathcal{F}_{pc}$ denote image classification head and patch classification head, respectively. For convenience, we denote the composite function $\mathcal{F}_b \circ \mathcal{F}_c$ as $\mathcal{F}_{cls}$ and $\mathcal{F}_b \circ \mathcal{F}_{pc}$ as $\mathcal{F}_{pcls}$. The image classification branch $\mathcal{F}_{cls}$ should enable the model to distinguish between CRC and healthy samples, so the classification loss is defined as follows:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}\left(p, y\right), p = \mathcal{F}_{cls}\left(I\right), \tag{1}$$

where I is the input image, y is the image-level label, p is the probability of lesion, and $\mathcal{L}_{CE}$ is the cross entropy (CE) loss function.

2.1 Loose Location Self-supervision Constraint

Loose Self-supervision Mechanism. To minimize the dependence on fine labeling, we devise a loose self-supervision mechanism that provides constraint information for accurate lesion localization. As shown in Figure 1, the lesion location remain close across adjacent slices.By defining θ_x and θ_y as the deviation of X and Y coordinates between adjacent lesion slices and creating a bounding box with the maximum of them based on intermediate slice, all lesion area will be within this box, leaving the background area healthy. Therefore, the method provides supervision information by ensuring background consistency.

On the temporal level, considering a sequence of CRC non-contrast CT images $\{I_k^n\}_{k=1}^K$, where I_k^n denotes the non-contrast CT image at position k. Patch features extracted by $\mathcal{F}_b$ are processed with $\mathcal{F}_{pc}$ to produce grid patch $p_k^n = \mathcal{F}_{pcls}\left(I_k^n\right)$. Here, $p_k^{n,i,j}$ denotes the probability of lesion for a patch in row i and column j with $i \in [1, h], j \in [1, w]$, and $h \times w$ is the total number of patches. To maintain background consistency, M_k^n is utilized to mask the area within the bounding box centered on the highest predicted lesion probability location (a, b) in p_k^n. Then the temporal consistency loss is defined as follows:

$$\begin{aligned} \mathcal{L}_t &= \sum_{k=2}^{K-1} \left(JS\left(P_k^n \| P_{k-1}^n\right) + JS\left(P_k^n || P_{k+1}^n\right)\right), P_k^n = p_k^n * M_k^n, \\ M_k^{n,i,j} &= 1 - \mathbf{1}(|i-a| \le \theta_x \text{ and } |j-b| \le \theta_y), (a,b) = \arg\max_{i,j} p_k^{n,i,j}, \end{aligned} \tag{2}$$

where $\mathbf{1}(*)$ is the indicator function, and P_k^n is the lesion masked grid patch at position k. The predicted distribution differences between the intermediate grid patch and both the previous and the latter are quantified respectively by $JS\left(P_k^n \| P_{k-1}^n\right)$ and $JS\left(P_k^n \| P_{k+1}^n\right)$ using Jensen-Shannon divergence [12]. This symmetric and bounded measure enhances the model's stability and robustness.

On the modality level, similar to temporal proximity, lesion locations across different modalities are close, leading to apply a similar loose location constraint. Considering a CRC arterial phase CT sequence $\{I_k^a\}_{k=1}^K$ and a CRC venous phase CT sequence $\{I_k^v\}_{k=1}^K$ of contrast CT, we calculate the grid patch probability

$p_k^a = \mathcal{F}_{pcls}\left(I_k^a\right)$ and $p_k^v = \mathcal{F}_{pcls}\left(I_k^v\right)$ respectively. Thus, the modality consistency loss is optimized by minimizing the background difference:

$$\begin{aligned} \mathcal{L}_m &= \sum_{k=1}^{K} \left(JS\left(P_k^n \| P_k^a\right) + JS\left(P_k^n || P_k^v\right)\right), \\ P_k^a &= p_k^a * M_k^n, P_k^v = p_k^v * M_k^n. \end{aligned} \tag{3}$$

Weakly Supervised Guidance. Loose self-supervision mechanism provides a certain degree of constraint information, yet effectively correcting initial misclassifications is challenging. In order to get better initial classification guidance, we use a small number of supervised samples (less than 10% of the total) and implement patch classification loss:

$$\mathcal{L}_{patch} = \mathcal{L}_{CE}\left(p^{n0}, 0\right) + \mathcal{L}_{CE}\left(p^{n1}, 1\right) + \mathcal{L}_{fuzzy}\left(p^{n1}\right), \tag{4}$$

where p^{n0} and p^{n1} are the grid patch prediction probability of healthy and CRC samples in non-contrast CT, respectively. $\mathcal{L}_{CE}$ is employed for optimizing few labeled CRC and healthy samples. $\mathcal{L}_{fuzzy}$ is the fuzzy classification loss [5] and used for unlabeled CRC samples. Therefore, location constraint significantly improves the accurate localization of lesion with weak supervision guidance.

Generally, the loose location self-supervision includes temporal consistency $\mathcal{L}_t$, modality consistency $\mathcal{L}_m$ and patch weakly supervised guidance $\mathcal{L}_{patch}$:

$$\mathcal{L}_{con} = \mathcal{L}_t + \mathcal{L}_m + \mathcal{L}_{patch}. \tag{5}$$

2.2 Mask Correction Loopback Mechanism

Further, a lesion location mask correction loopback mechanism is devised to enhance the reliability of the prediction results, which constrains the strong dependence between category label and lesion location.

Specifically, we upsample lesion masked grid patch P^n of the non-contrast CT image I^n to match the input size and use it to mask image. The lesion masked image is then input back into the classification branch, and it is expected to be classified as healthy. So the loopback consistency loss is defined as follows:

$$\mathcal{L}_{loop} = \mathcal{L}_{CE}\left(\mathcal{F}_{cls}\left(I^n * \mathcal{U}\left(P^n\right)\right), 0\right), \tag{6}$$

where $\mathcal{U}\left(p^n\right)$ is a mapping function that upsamples P^n. $I^n * \mathcal{U}\left(p^n\right)$ represents the lesion masked image. The above loss function is only for CRC samples.

The loopback consistency loss continuously optimizes the model by masking the lesion location and binds localization and overall classification results to strengthen the consistency and increase the reliability of the model.

2.3 Complete Framework

The combination of loose lesion location self-supervision and fuzzy classification loss allows accurate lesion localization without extensive fine labeling. The

Table 1. Slice-level classification and localization performance. 'P-pre.', 'P-rec.' and 'P-IoU' denote patch precision, patch recall, and patch IoU, respectively.

Method		Classification					Localization		
		Accuracy	Precision	Recall	Specificity	F1-score	P-pre.	P-rec.	P-IoU
Unsupervised	GradCAM [20]			/			47.70%	62.58%	37.12%
	EigenCAM [16]			/			58.54%	78.67%	50.42%
Fully supervised	Sahoo [19]	89.36%	81.19%	84.17%	91.60%	82.65%	69.34%	86. 41%	62.52%
	ResNet50 [7]	88.53%	78.60%	85.08%	90.02%	81.71%	61.07%	88.52%	56.59%
	MobileNetV3-L [9]	87.86%	78.73%	81.75%	90.49%	80.21%	62.15%	87.77%	57.20%
	RegNetY-128G [18]	89.16%	79.45%	86.33%	90.38%	82.75%	61.11%	81.06%	53.48%
	EfficientNetV2-L [23]	87.36%	74.72%	87.67%	87.22%	80.67%	67.25%	85.29%	60.26%
	ConvNeXt-L [14]	90.29%	82.08%	86.67%	91.85%	84.31%	71.15%	82.10%	61.60%
	ConvNeXtV2-L [27]	90.44%	83.10%	85.67%	92.50%	84.37%	65.81%	**90.85%**	61.72%
8.3% supervised	Ours	**91.19%**	**83.72%**	**87.83%**	**92.64%**	**85.73%**	**79.19%**	75.32%	**62.87%**

mask correction loopback mechanism constrains the strong dependence between classification and localization, enhancing the reliability of CRC diagnosis.

We train our model in two stages. Firstly, the image classification branch is trained by optimizing $\mathcal{L}_{cls}$ to enable the model to distinguish between CRC and healthy samples. Secondly, $\mathcal{L}_{cls}$, $\mathcal{L}_{con}$ and $\mathcal{L}_{loop}$ are jointly optimized to train both image classification branch and patch classification branch with the following loss function:

$$\mathcal{L} = \alpha\mathcal{L}_{cls} + \beta\mathcal{L}_{con} + \gamma\mathcal{L}_{loop}, \tag{7}$$

where α, β and γ are balance parameters.

3 Experiments and Results

3.1 Dataset

In this paper, we collect colorectal CT scan dataset from three medical centers and each patient contains three modalities: non-contrast, arterial and venous, of which arterial and venous are contrast CT. A total of 3,178 CRC patients and 887 healthy controls are included. Please find the details of the dataset in the *supplementary material*. In our experiment, the training set includes three modalities data of 3,028 CRC and 737 healthy controls (75,642 lesion slices and 64,182 healthy slices with equal number of each modality). We only use 6,304 non-contrast CT slices (8.3% of the total) with lesion location bounding box labeling to provide supervision information. The test set consisted of non-contrast CT scan from 150 CRC patients and 150 healthy controls (1,200 lesion slices and 2,786 healthy slices) is used to evaluate our method's effectiveness.

3.2 Implementation Details

The network's backbone $\mathcal{F}_b$ is the ConvNeXtV2-L [27] pretrained on ImageNet [4] provided by timm library [26], and the last global adaptive pooling layer and the

Table 2. Patient-level classification performance.

Method	Accuracy	Precision	Recall	Specificity	F1-score
Sahoo [19]	84.33%	83.67%	85.33%	83.33%	84.49%
ResNet50 [7]	84.00%	82.28%	86.67%	81.33%	84.42%
MobileNetV3-L [9]	83.33%	82.89%	84.00%	82.67%	83.44%
RegNetY-128G [18]	84.00%	82.28%	86.67%	81.33%	84.42%
EfficientNetV2 [23]	83.33%	80.49%	88.00%	78.67%	84.08%
ConvNeXt-L [14]	85.33%	84.87%	86.00%	84.67%	85.43%
ConvNeXtV2-L [27]	85.67%	85.91%	85.33%	86.00%	85.62%
Ours	**87.00%**	**86.27%**	**88.00%**	**86.00%**	**87.13%**

fully connected layer are removed and regarded as image classification head $\mathcal{F}_c$. We adopt a three layers convolutional network for patch classification head $\mathcal{F}_{pc}$. Conv2d, BatchNorm2d and ReLU are successively added in the first two layers and the final layer is Conv2d for classifying each patch.

Initially, we only train the image classification branch for 10 epochs and then the entire network is trained for 10 epochs with deviation $\theta_x = 1, \theta_y = 1$ and balance parameters $\alpha = 0.5$, $\beta = 1$, $\gamma = 0.5$. The default batch size is 48. The optimizer is Adam [11], and the initial learning rate is $1e^{-4}$ with cosine annealing scheduler for each cycle, aligned with the overall training epochs. The size of input CT image is 224×224 px.

We will compare our method in slice-level classification and localization, as well as patient-level classification, with state-of-the-art (SOTA) methods.

3.3 Comparison with SOTA

Slice-level Experiment and Analysis. To validate our method, we compare classification and localization performance at slice level in Table 1. Grad-CAM [20] and EigenCAM [16] are two feature attribution methods, both utilizing the ConvNeXtV2-L backbone in experiments. Sahoo [19] is YOLOv8-based [24] detection method and supervised by fully bounding box annotations. ResNet50 [7], MobileNetV3-L [9], RegNetY-128G [18], EfficientNetV2-L [23], ConvNeXt-L [14] and ConvNeXtV2-L [27] are SOTA CNN-based image classification methods and we regard them as backbone, then add a location branch using fully annotations of three modalities to train.

It is evident that our method excels in slice-level classification tasks. This is mainly because the proposed loose location self-supervision and mask correction loopback mechanism, enhancing the model's ability to distinguish between CRC and healthy samples. For lesion localization, considering that our network is not designed for regression, direct comparison of IoU would be inappropriate. Therefore, we assess the patch IoU instead. Specifically, the original image is divided into grid patches, and we calculate the patch overlap rate for both predicted and actual boxes. So patch precision (P-pre.), patch recall (P-rec.) and patch IoU (P-IoU) are used to evaluate the localization performance. Our method achieves

Table 3. Ablation study on different loss terms.

Index	w/o $\mathcal{L}_{cls}$	w/o $\mathcal{L}_{con}$	w/o $\mathcal{L}_{loop}$	Ours
P-pre.	61.16%	18.59%	69.45%	**79.19%**
P-rec.	58.18%	16.71%	65.96%	**75.32%**
P-IoU	42.48%	9.65%	51.13%	**62.87%**

the best performance in terms of P-pre. and P-IoU but not excels in P-rec. due to lack of extensive fine labeling guidance. The results indicate that with only 8.3% annotations, our approach achieves promising results compared to feature attribution methods and those requiring massive annotations.

Patient-level Experiment and Analysis. Table 2 shows patient-level classification performance, which employs a threshold method for diagnosis by analyzing the entire colorectal scan sequence. If more than 7 slices are classified as lesion, it indicates CRC. Once again, our method demonstrates superior performance. For the whole scan sequence of patients in clinical diagnosis, the model can more effectively distinguish between lesion and healthy slices, further confirming our method's robustness and reliability.

Lesion Localization Visualization. Due to the specificity of medical diagnosis, the reliability is essential factor of the diagnostic method. A major advantage of our method is its ability to provide reliable predictions with minimal labeling. The proposed method predicts the lesion area by mapping the patch with the highest lesion probability onto the original CT image. The qualitative visual results of EigenCAM [16], ConvNeXtV2-L [27] with fully annotations (Fully) and Ours for lesion location are given in the *supplementary material*.

3.4 Ablation Study

To verify the effectiveness of each component, ablation study is conducted on three loss function terms: $\mathcal{L}_{cls}$, $\mathcal{L}_{con}$, and $\mathcal{L}_{loop}$. Table 3 gives the quantitative evaluation of lesion localization performance, showing that all loss terms have contributed to the final result. Notably, $\mathcal{L}_{cls}$ can maintain the discriminative capability of model to distinguish between CRC and healthy samples and lacking it leads to locating performance degradation. Due to the lack of the loose location self-supervision constraint, abandoning $\mathcal{L}_{con}$ significantly impairs the model's ability of accurate lesion localization. The absence of $\mathcal{L}_{loop}$ shows that the mask correction loop mechanism further enhances model performance. Therefore, the three terms work together to improve lesion localization performance.

4 Conclusion

In this paper, we propose loose lesion location self-supervision enhanced CRC diagnosis, which is composed of loose location self-supervision constraint and

mask correction loopback. Without the need for extensive lesion location labeling, we apply location consistency constraint to accurately localize lesion from both temporal and modality perspectives. Masking the lesion location enhances the consistency between category label and lesion location, ensuring reliable diagnostic. Extensive experiments show that our method achieves accurate classification and provides reliable lesion localization, offering an effective diagnostic tool for CRC. In the future, we will focus on the extension of self-supervision and mask correction loopback mechanism into more general classification tasks and applying the proposed method to auxiliary diagnosis in clinical practice. Furthermore, we will also devote ourselves to considering additional criteria for diseases with significant location variations to avoid erroneous constraints.

Acknowledgments. This work is supported by National Natural Science Foundation of China (62376248), and the Huadong Medicine Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMZ24H160001.

Disclosure of Interests. The authors declare no competing interests.

References

1. Akilandeswari, A., Sungeetha, D., Joseph, C., Thaiyalnayaki, K., Baskaran, K., Jothi Ramalingam, R., Al-Lohedan, H., Al-Dhayan, D.M., Karnan, M., Meansbo Hadish, K., et al.: Automatic detection and segmentation of colorectal cancer with deep residual convolutional neural network. Evidence-Based Complementary and Alternative Medicine **2022** (2022)
2. Argilés, G., Tabernero, J., Labianca, R., Hochhauser, D., Salazar, R., Iveson, T., Laurent-Puig, P., Quirke, P., Yoshino, T., Taieb, J., et al.: Localised colon cancer: Esmo clinical practice guidelines for diagnosis, treatment and follow-up. Annals of Oncology **31**(10), 1291–1305 (2020)
3. Chan, H.P., Samala, R.K., Hadjiiski, L.M., Zhou, C.: Deep learning in medical image analysis. Deep Learning in Medical Image Analysis: Challenges and Applications pp. 3–21 (2020)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
5. Feng, Z., He, Y., Wang, X., Gao, X., Lei, J., Jin, C., Song, M.: One-sample guided object representation disassembling. Advances in Neural Information Processing Systems **33**, 21878–21888 (2020)
6. Han, Y.E., Cho, Y., Park, B.J., Kim, M.J., Sim, K.C., Sung, D.J., Han, N.Y., Lee, J., Park, Y.S., Yeom, S.K., et al.: Development and multicenter validation of deep convolutional neural network–based detection of colorectal cancer on abdominal ct. European Radiology pp. 1–11 (2024)
7. 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)

8. Hicham, K., Laghmati, S., Hamida, S., El Ghazi, A., Tmiri, A., Cherradi, B.: Assessing the performance of deep learning models for colon polyp classification using computed tomography scans. In: 2023 3rd International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET). pp. 01–06. IEEE (2023)
9. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1314–1324 (2019)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (1991)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2980–2988 (2017)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
15. Lu, L., Dercle, L., Zhao, B., Schwartz, L.H.: Deep learning for the prediction of early on-treatment response in metastatic colorectal cancer from serial medical imaging. Nature communications **12**(1), 6654 (2021)
16. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
17. Pacal, I., Karaboga, D., Basturk, A., Akay, B., Nalbantoglu, U.: A comprehensive review of deep learning in colon cancer. Computers in Biology and Medicine **126**, 104003 (2020)
18. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10428–10436 (2020)
19. Sahoo, P.K., Gupta, P., Lai, Y.C., Chiang, S.F., You, J.F., Onthoni, D.D., Chern, Y.J.: Localization of colorectal cancer lesions in contrast-computed tomography images via a deep learning approach. Bioengineering **10**(8), 972 (2023)
20. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
21. Shaukat, A., Levin, T.R.: Current and future colorectal cancer screening strategies. Nature Reviews Gastroenterology & Hepatology **19**(8), 521–531 (2022)
22. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. Ca Cancer J Clin **73**(1), 17–48 (2023)
23. Tan, M., Le, Q.: EfficientNetV2: Smaller models and faster training. In: International conference on machine learning. pp. 10096–10106. PMLR (2021)
24. Terven, J., Cordova-Esparza, D.: A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond. arXiv preprint arXiv:2304.00501 (2023)

25. Uemura, T., Näppi, J.J., Hironaka, T., Kim, H., Yoshida, H.: Comparative performance of 3D-DenseNet, 3D-ResNet, and 3D-VGG models in polyp detection for ct colonography. In: Medical Imaging 2020: Computer-Aided Diagnosis. vol. 11314, pp. 736–741. SPIE (2020)
26. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019). https://doi.org/10.5281/zenodo.4414861
27. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16133–16142 (2023)
28. Yao, L., Li, S., Tao, Q., Mao, Y., Dong, J., Lu, C., Han, C., Qiu, B., Huang, Y., Huang, X., et al.: Deep learning for colorectal cancer detection in contrast-enhanced ct without bowel preparation: A retrospective, multicentre study. http://dx.doi.org/10.2139/ssrn.4617045 (2023)
29. Yao, L., Xia, Y., Zhang, H., Yao, J., Jin, D., Qiu, B., Zhang, Y., Li, S., Liang, Y., Hua, X.S., et al.: Deepcrc: Colorectum and colorectal cancer segmentation in ct scans via deep colorectal coordinate transform. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 564–573. Springer (2022)
30. Yin, Z., Yao, C., Zhang, L., Qi, S.: Application of artificial intelligence in diagnosis and treatment of colorectal cancer: A novel prospect. Frontiers in Medicine **10**, 1128084 (2023)