MULTI-VIEW NETWORK FOR COLORECTAL POLYPS DETECTION IN CT COLONOGRAPHY

Mohamed Yousuf [⋆‡] *Samir Harb* [⋆‡‡] *Islam Alkabbany* [⋆◇] *Asem Ali* [⋆] *Salwa Elshazley* [†] *Aly Farag*[⋆]

[⋆] Computer Vision and Image Processing Laboratory (CVIP), University of Louisville, Louisville, KY
[‡] Faculty of Engineering, Ain Shams University, Cairo, Egypt
[‡‡] Higher Technological Institute, 10th of Ramadan City, Egypt
[◇] Faculty of Engineering, Assiut University, Assiut, Egypt
[†] Kentucky Imaging Technologies, Louisville, KY

ABSTRACT

Early diagnosis of colorectal polyps, before they turn into cancer, is one of the main keys to treatment. In this work, we propose a framework to help radiologists in reading CT scans and identifying candidate CT slices that have polyps. We propose a colorectal polyps detection approach which consists of two cascaded stages. In the first stage, a CNN-based model is trained and validated to detect polyps in axial CT slices. To narrow down the effective receptive field of the detector neurons, the colon regions are segmented and then fed into the network instead of the original CT slice. This drastically improves the detection and localization results, e.g., the mAP is increased by 36%. To reduce the false positives generated by the detector, in the second stage, we propose a multi-view network (MVN) that classifies polyp candidates. The proposed MVN classifier is trained using sagittal and coronal views corresponding to the detected axial views. The approach is tested in 50 CTC-annotated cases, and the experimental results confirm that after the classification stage, polyps can be detected with an AUC $\sim 95.27\%$.

***Index Terms*—** Colorectal cancer, colon polyp, computerized tomography (CT), polyp Detection, CNN, segmentation.

1. INTRODUCTION

Colorectal cancer (CRC) begins as small growths (polyps) that attach to the luminal wall of the colon or rectum, which must be diagnosed and treated promptly. Optical colonoscopy (OC), a procedure in which the colon and rectum are viewed through a lighted flexible tube with a camera at the end, is the standard screening approach. However, OC is an expensive and invasive procedure. On the other hand, Computed Tomographic Colonography (CTC), a screening method in which radiologists detect colon polyps from CTC images, is a noninvasive, inexpensive, and gives clinically acceptable performance [1–3]. Chini et al., [4] showed that CTC is an acceptable alternative to OC for polyp size ≥ 5 mm. Furthermore, compared to other noninvasive CRC screening methods, the precision of CTC is much higher for sensitivity and specificity [1, 5]. We propose a polyps detection approach to improve the performance of CTC screening. The proposed approach will help radiologists in reading CT scans and identifying candidate CT slices that have polyps and then referring to polyp locations.

THIS WORK HAS BEEN FUNDED BY NSF GRANT 2124316.

Table 1. Summary of Polyp Analysis using CT Scans.

Approach	# Scans	$\mathcal{S}\%$	Limitations*
Shape-based model [6]	10	71	shape variant
Texture and shape analysis [7]	56	100	
Shape & size invariant [8]	249	95	single type
DETALE transfer learning [9]	154	94	polyp or non-polyp only
Deep ensemble CNN [10]	403	90.5	
DCNN-CADe [11]	144 polyps	93	
2D CNN [12]	N/A	97	
3D-GLCM CNN [13]	63 polyps	90.1	
3D-Dense CNN [14]	403	95.1	categorization only
Machine learning [15]	106	82	
Seg + 3D CNN [16]	107	80	
MM-GLCM CNN [17]	100	94.5	

* None of the approaches detects the location and the size of a polyp

Over the past two decades many researchers exploited computer vision algorithms to automatically detect or classify colonic polyps. Table 1 shows a summary of the literature review for algorithms that perform polyp analysis from CT scans. The summary categorizes these algorithms and shows the number of used CT scans as well as the reported sensitivity $\mathcal{S}$, for each algorithm. Although the reported sensitivities, in Table 1, are promising, there are many limitations and drawbacks in these approaches. The approaches in [6, 7] are shape-based algorithms but polyps come in a variety of shapes, sizes, and types. Wijk et al. [8] focused on one polyp type with lots of false positives. Furthermore, [9–13] classify polyp from non-polyp candidates and [14–16] are focusing on classifying the polyp category (Adenomatous, Serrated, Inflammatory polyps). Unfortunately, there is no baseline dataset so the authors, of each approach in Table 1, used their private dataset for evaluation. In this work, experiments are conducted on the dataset that was used by Zhang et al., [17]. However, Zhang et al., [17] focused on distinguishing malignant from benign lesions, but we focus on polyps detection and localization. The only public dataset that has CT scans with colonic polyps annotation is presented by The Cancer Imaging Archive (TCIA), ACRIN-6664 [18]. However, only the slice number, which has polyps, is provided not the location and size of each polyp within the slice. Finally, none of the approaches, in Table 1, detects the exact location and size of a polyp within a CT scan except Grosu et al., approach, [15], which detects polyps before classification, but with a detection rate 73%.

The proposed framework aims to identify the accurate location and size of potential polyps by locating the slice and the bounding box for the detected polyp. Colonic polyp detection is a challenging problem due to the uncertainties in acquiring CT scans, e.g., preparation artifacts, polyp size, and location [19]. To develop an efficient approach that overcomes polyps detection and localization challenges, the proposed algorithm focuses on colon regions. So, a preprocessing step is performed to segment colon regions. Then, the segmented regions are fed into a CNN-based detector for polyps detection. The detector results are fed into the proposed CNN-MVN module to enhance the final results. The main contributions of the proposed work include developing:

1. a CNN-based polyp detector that is trained and validated using axial CT scans,
2. a colon segmentation approach to guide the proposed detectors focusing on the colon wall, and
3. a multi-view fusion network (MVN) classifies polyp candidates generated by the proposed detector. The proposed MVN classifier is trained using sagittal and coronal views corresponding to the detected axial views.

2. PROPOSED APPROACH

The proposed approach for colonic polyps detection and localization uses segmented axial CT slices, which are fed into a CNN model to localize candidate polyps, as shown in Fig. 1. A segmentation step is used to guide the CNN detector to the regions of interest only, the colon wall. This changes the neuron's effective receptive field to focus on colon regions. However, the detector may generate false positives. To eliminate these false positives, we exploit the other two views of a CT scan (i.e., a CT scan is a volume in DICOM format and other views can be projected not only the axial view). Therefore, we train a classifier using the three views, as shown in Fig. 3. The proposed MVN classifier uses three 2D images (sagittal, coronal, and axial views) for each candidate generated by the detectors. Moreover, the proposed classifier avoids the drawbacks of using a volume in the training process (e.g., 3D-CNN [14]) and consumes less time and memory compared to the 3D network.

2.1. Colon Segmentation

Colon segmentation is a challenging problem due to the colon's asymmetric topology. Also, uncertainties appear due to the presence of Hounsfield (HU) intensity regions consisting of air, soft tissue, and high-attenuation structures like the bones. In addition, complications result due to the presence of residual stool, parts of the diaphragm, lungs, and disconnected colon segments [20]. In this paper, we propose a segmentation approach that involves multiple steps, as shown in Fig. 2. The first step is to calculate the empirical distributions of Hounsfield intensities in a DICOM volume, as shown in Fig. 2-b. The main components of a colon are the air, for which the characteristic peaks are almost at -1000 HU [21], and the fluid whose Hounsfield intensity is greater than 300 HU.

To extract the colon components, first, we estimate the marginal densities of air, fat, muscle, and fluid by fitting four Gaussian components using the Expectation Maximization (EM) algorithm, as shown in Fig. 2-b. Then, we identify colon regions using two thresholds. The first threshold t_1 is between air and fat, and the second threshold t_2 is between muscle and fluid. The HU values of colon regions should be $< t_1$ and $> t_2$. However, this simple thresholding technique cannot isolate colon regions from non-colon regions. Therefore, we use this initial segmentation to extract the rectum region, which can be easily identified as a disk-like region that has a low HU, in the first part of the DICOM volume, as shown in Fig. 2-d. This region is used as a starting seed, from which other colon regions are extracted by region growing. However, since there are non-colon parts that are interwind with colon parts, this yields errors in the region growing step. Therefore, restricted region growing is performed using the morphological operation to guarantee more separation between the colon and non-colon classes. The output of the region growing step, Fig. 2-e, is used as a seed for Graph Cut approach [22] to generate the final segmentation Fig. 2-f.

2.2. CNN-based polyps detection approach

After performing the segmentation step, we use the segmentation as a mask for colon regions since the detection algorithm would focus only on the colon regions. Therefore, the original CT slice is multiplied by this binary mask, as shown in Fig. 1. To make sure that the colon wall is included in the segmented regions, the segmented regions are dilated before the multiplication. The output of this step, as shown in Fig. 1, is the masked axial view, which is fed into the YOLO model [23]. The model is trained to extract all the true polyps from the axial view, even with a high number of false positives. Therefore, the confidence score threshold is chosen to have a low value to produce all possible true positive regions.

2.3. Multi-view fusion network (MVN)

Since the detector produces many false positives from the axial view, an algorithm is needed to filter these false positives. We tried to mimic what radiologists do when they find a potential polyp candidate. They usually confirm or discard this finding by examining the other two views of the polyp candidate (i.e., sagittal and coronal views). This is why we extract the corresponding sagittal and coronal views to be fed into our proposed MVN model. Our hypothesis is based on that most polyps have the shape of a small protruding mound and they should appear in the three views as a small protruding mound, unlike the non-polyp areas, e.g. colon folds. Figure 4 shows examples of different false positive samples. The top row illustrates the axial views, and each has a geometric feature similar to polyps, so they can deceive the detector. However, the sagittal and coronal views, in the bottom row, show that the geometric and appearance features of the polyp are not presented. Therefore, using the other views will filter the detector results. However, in some hard cases (e.g. Fig. 4-e) the three views show the presence of the geometric feature of a polyp while it is not a true polyp.

To integrate information from the different views, inspired by Markov chain model [24], we assume that the polyp classification is conditionally independent due to the different views. Consequently, the joint probability over all views factorizes into the conditional probabilities over the separate views, using the chain rule, as follows.

$$P(Y|X) = P(Y_c|X, Y_s, Y_a)P(Y_s|X, Y_a)P(Y_a|X), \quad (1)$$

where $X = \{X_c, X_s, X_a\}$ is the input sequence of the three views and $Y = \{Y_c, Y_s, Y_a\}$ is the output sequence, which should be predicted. As shown in Fig. 3, the learned features $\{f_c, f_s, f_a\}$ are extracted from the three views $\{X_c, X_s, X_a\}$, respectively. According to Eq. 1, estimating axial prediction Y_a depends only on the extracted features from axial slice, so its conditional probability is approximated as follows.

$$P(Y_a|X) = Sigmoid(FC(f_a)), \quad (2)$$

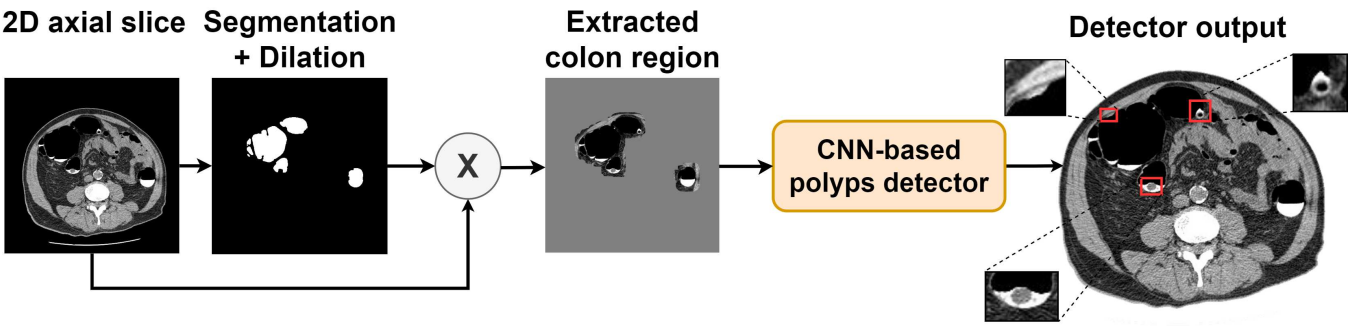

Fig. 1. The first stage of the proposed CT-based approach. An input axial CT slice is segmented. Then it is fed into a CNN-based detector to localize polyp candidates. The output shows the location and the size of any potential polyp that needs to be verified.

where $FC(.)$ is a fully connected network. On the other hand, estimating the sagittal prediction Y_s depends on the axial prediction Y_a as well as the extracted features from both sagittal and axial views so its conditional probability can be approximated as follows.

$$P(Y_s|X, Y_a) = Sigmoid(FC([f_s, h_a, y_a])). \quad (3)$$

Finally, estimating the coronal prediction Y_c depends on both axial and sagittal predictions as well as the extracted features from all views, so its conditional probability can be approximated as follows.

$$P(Y_c|X, Y_a, Y_s) = Sigmoid(FC([f_c, h_s, y_s, y_a])). \quad (4)$$

This model is trained to check if the detected candidate is a polyp or not in the three corresponding views of the candidate. The final score reflects the score of all three views to obtain the best classification results.

3. EXPERIMENTAL RESULTS

3.1. Dataset

The dataset used to train and validate the proposed modules, was provided by CTC experts from the University of Wisconsin. The data consists of scans, in the supine and prone positions, for 50 patients. The scans have 59 annotated polyps larger than 6 mm. The annotation of the lesions was done by one of three experienced radiologists. The train-validation ratio is 80/20 while cross-validation is used due to the limited number of scans in the dataset.

3.2. Training and testing procedure for detector stage

To evaluate the detection model, first, the proposed colon segmentation approach has been applied to the input axial view of CT slices to be fed into the next stage for detecting polyps. Next, data augmentation was used to make different variations to the dataset by flipping some of the images horizontally and applying different exposure, saturation, and brightness. These augmented and segmented images have been used to train the model. To present the importance of the segmentation step on axial view images, a YOLO model has been trained with the original 2D CT slices without the segmentation step. This model gives only 44% mean Average Precision (mAP).

Also, for the sake of comparison, masked slices have been used to train different models e.g., YOLO-V5 [25], YOLO-V7 [26], Faster-RCNN with Resnet-50 [27], and Faster-RCNN with Resnet-101 [27]. As shown in Table 2, the YOLO-V7 detector has the highest mean Average Precision (mAP), so it has been chosen to become the detector of the first stage of the proposed framework. As expected YOLOv7 outperforms YOLOv5 since it is a bigger model and can fit more complex mapping. On the other hand, Dynamic RCNN with ResNet 50 [28], Retina Net with Efficient Net backbone [29], Sparse RCNN [30], and Swin Transformer [31] were implemented and trained for the polyp detection task. However, their results were not included in Table 2 as these models performed poorly, with a very low mAP. These models were unable to detect very small objects such as polyps in 2D slices, as a polyp in a 2D CT slice would only occupy on average $4 - 5\%$ of the total image size. Note that we use the mean Average Precision rather than the mean Average Recall (mAR), or the F1 score since our main concern is not to miss any polyp rather than giving some false positives within the detected regions. All the detectors training's batch size is 16, but all other hyperparameters were the original hyperparameters from each model's original papers. All the training was performed on Nvidia TITAN RTX 24 Gb. Moreover, all models were trained to a high number of epochs (2000 epochs) to get the best validation score epoch.

Table 2. The results of the CT-based detection stage.

Model	Sensitivity	mAP
YOLO-V5 [25]	86.67 %	79.7 %
YOLO-V7 [26]	88.05%	85.7 %
Faster RCNN R-50 [27]	69.1%	71 %
Faster RCNN R-101 [27]	70.6%	71.7 %

3.3. Training and testing procedure for classifier stage

To reduce the false positives generated by the detector, the proposed multi-view fusion network, Fig. 3, is trained using sagittal and coronal views corresponding to the detected axial views, to classify the candidates. Additional experiments have been conducted to investigate more classification architectures compared to the proposed MVN model. In the first experiment, instead of using the detected 2D axial view and its corresponding sagittal and coronal views, i.e., three 70×70 images, the corresponding volume ($70 \times 70 \times 15$) is extracted from the DICOM. Then, the 3D-CNN [14] model has been trained using the extracted volumes. The 3D-CNN model, as expected, needs high computational power to be trained and it needs a larger dataset for the training. As shown in Table 3, it is almost 7 times the size of the 2D-CNN model as shown in the number of parameters.

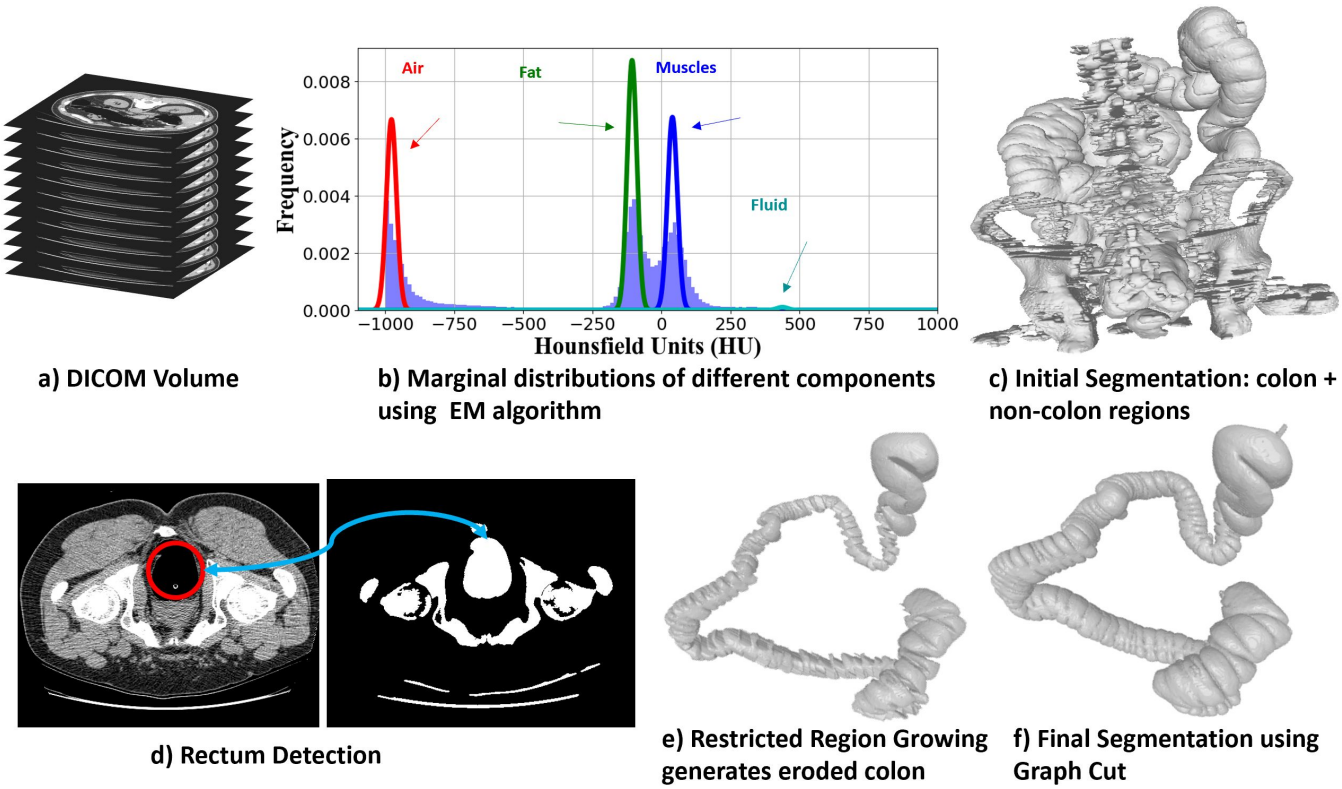

Fig. 2. The Proposed Colon Segmentation Approach

In the second experiment, instead of using each of the three corresponding views as a separate image, the three views were combined as three channels, then the three channels images were used to train the depthwise convolution-based model [32]. The depthwise convolution layer had been used in this experiment because the spatial relations are different in the three views, so, the standard CNN architecture cannot be used.

In the third experiment, we use some of the most successful classifiers to compare them with the previous results as shown in Table 3. The implemented classifiers are: ResNet 18 [33], VGG-19 [34], GoogleNet [35], DenseNet [36], SENet [37], PNASNet [38], MobileNetV2 [39], ResNext [40], and ShuffleNetV2 [41]. As shown in Table 3, the proposed multi-view fusion network (MVN) has the highest number in the sensitivity 94.80% and area under the curve 95.27%, indicating that the classifier has rejected most false positive candidates that had been extracted in the previous detector stage, and most of the true positive results are correctly classified. We believe that our proposed MVN model outperforms other models because the model has been customized to fit perfectly on the polyp classification task with the lowest number of parameters needed, which leads to being the fastest model while performing the testing procedure. All the classifiers training's batch size = 256, epochs= 2000 with early stopping, and the training was performed on Nvidia TITAN RTX 24 Gb. All models have been trained from scratch.

To illustrate the effect of each stage in the proposed approach, an experiment has been conducted using 19 2D axial slices, each slice having a polyp. The standard YOLO detector missed 11 polyps, i.e., $FN = 11$, in addition to three false positive samples. So, we reduced its confidence score threshold, as in the previous experiments, to detect all the polyps. Although we make $FN = 0$, this configuration extracts $FP = 13$ false positive samples. After applying the proposed classifier to these candidates, these false positive samples have been decreased to only one sample.

Table 3. The results of the MVN classification stage.

Model	Parameters	Sensitivity	AUC
proposed MVN	520,067	96.92%	95.27%
3D CNN [14]	1,296,577	87.80%	93.52%
Depthwise conv [32]	399,314	84.73%	89.97%
ResNet-18 [33]	11,186,274	79.06%	84.90%
VGG-19 [34]	20,036,418	82.59%	86.31%
GoogleNet [35]	6,158,050	80.58%	85.12%
DenseNet [36]	6,956,298	83.85%	90.53%
SENet [37]	11,256,250	81.46%	88.94%
PNASNet [38]	450,594	79.57%	86.82%
MobileNetV2 [39]	2,286,674	79.82%	85.71%
ResNext [40]	9,120,578	78.56%	82.53%
ShuffleNetV2 [41]	1,263,854	79.31%	84.22%

We note that the proposed MVN model may fail in hard conditions, which have geometric and appearance features similar to polyp in all of the three views. As shown in Fig. 4, cases (a) to (d) have been correctly classified because at least one of the views does not have polyp-like features, but as shown in case (e), the classifier identifies the sample as a polyp because its three views have polyp-like features, although it is not a polyp.

4. CONCLUSION

In this work, we proposed a colorectal polyp detector to identify the locations and sizes of the polyps for CT colonography. The proposed

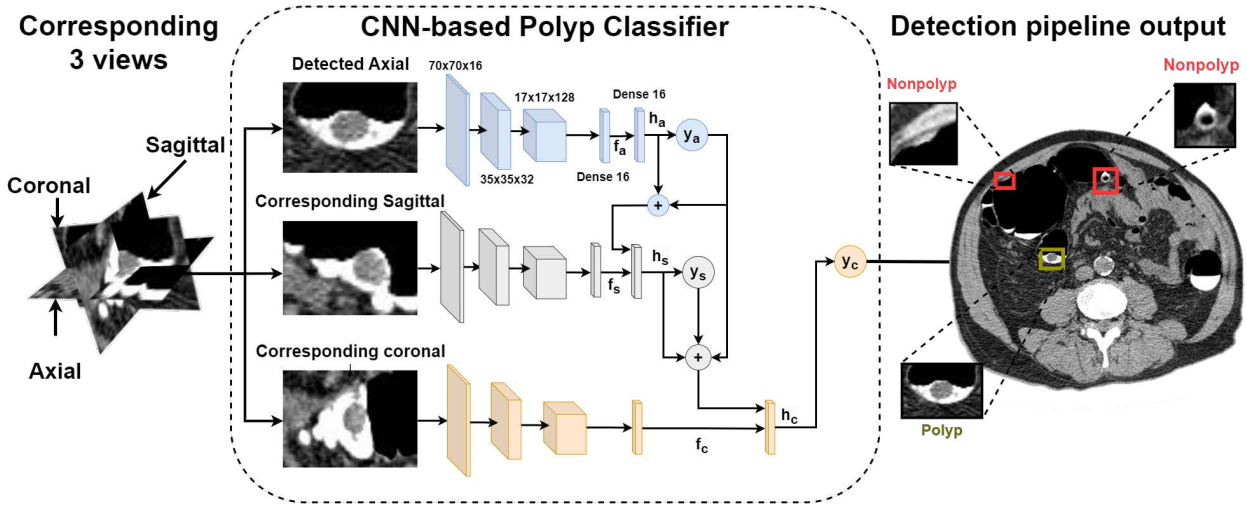

Fig. 3. Multi-view fusion network (MVN) is the second stage of the proposed CT-based approach. The detected axial candidate and its corresponding sagittal and coronal views are fed into a CNN-based classifier to classify a candidate using a Markov chain network. The output shows the location and the size of the final detected polyp after removing false positives.

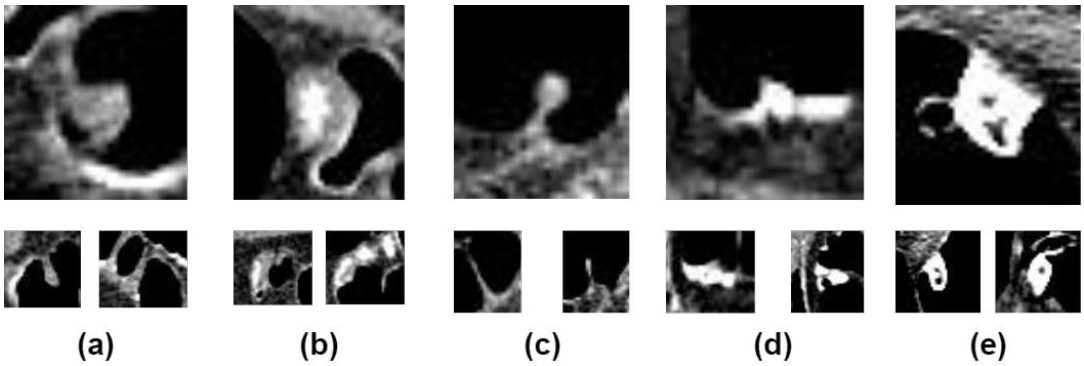

Fig. 4. Different false positive samples. The top is the axial view showing the geometric appearance and features of polyps. (a-d) false positives that do not have consistent features in the three views. In hard cases (e.g., e) the three views confirm the presence of the geometric feature of a polyp, so the proposed approach may fail.

approach consisted of two cascade stages: a detector, to detect polyp candidates, followed by a classifier, to filter the detected candidates. The proposed CNN-based detector was guided by changing the neuron's effective receptive field, using the segmented colon wall, which drastically enhanced the detection performance. Since the detector produced false positives, each detected candidate was fed into the proposed multi-view fusion network (MVN) classifier to exploit the different views of the CT scans for each candidate. This step refined the detection by rejecting false positives. The high sensitivity $\sim 94.8\%$ and AUC $\sim 95.3\%$ of the proposed approach illustrate that it can be adopted by radiologists to read CT scans in a short time.

5. COMPLIANCE WITH ETHICAL STANDARD

This study was carried out in accordance with the principles of the Declaration of Helsinki. The approval was granted by the Ethics Committee of the University of Wisconsin.

References

[1] C D Johnson, M Chen, A Y Toledano, J P Heiken, A Dachman, M D Kuo, C O Menias, B Siewert, J I Cheema, R G Obregon, et al., "Accuracy of ct colonography for detection of large adenomas and cancers," *NEJM*.

[2] Cesare Hassan and Perry J Pickhardt, "Cost-effectiveness of ct colonography," *Radiologic Clinics*, 2013.

[3] P. J Pickhardt, "Ct colonography for population screening: ready for prime time?," *Digestive diseases and sciences*, 2015.

[4] A. Chini, M Manigrasso, G Cantore, R Maione, M Milone, F Maione, and G D De Palma, "Can computed tomography colonography replace optical colonoscopy in detecting colorectal lesions?: State of the art," *Clinical Endoscopy*, 2022.

[5] P J Pickhardt, P M Graffy, B Weigman, N Deiss-Yehiely, C Hassan, and J M Weiss, "Diagnostic performance of multitarget stool dna and ct colonography for noninvasive colorectal cancer screening," *Radiology*, 2020.

[6] R M Summers, C F Beaulieu, L M Pusanik, J D Malley, R B Jeffrey Jr, D I Glazer, and S Napel, "Automated polyp detector for ct colonography: feasibility study," *Radiology*, 2000.

[7] W Hong and F Qiu, "A pipeline for computer aided polyp detection," *IEEE Trans Vis Comput Graph*, 2006.

[8] V Wijk, V Ravesteijn, Frank M Vos, R Truyen, A Vries, J Stoker, and L Vliet, "Detection of protrusions in curved folded surfaces applied to automated polyp detection in ct colonography," in *MICCAI*. Springer, 2006.

[9] J Näppi, T Hironaka, D Regge, and H Yoshida, "Deep transfer learning of virtual endoluminal views for the detection of polyps in ct colonography," in *Medical imaging 2016: computer-aided diagnosis*. SPIE.

[10] K Umehara, J Näppi, T Hironaka, D Regge, T Ishida, and H Yoshida, "Deep ensemble learning of virtual endoluminal views for polyp detection in ct colonography," in *Medical Imaging: Computer-Aided Diagnosis*, 2017.

[11] J Näppi, P Pickhardt, D Kim, T Hironaka, and H Yoshida, "Deep learning of contrast-coated serrated polyps for computer-aided detection in ct colonography," in *Medical Imaging 2017: Computer-Aided Diagnosis*.

[12] Y Chen, Y Ren, L Fu, J Xiong, R Larsson, X Xu, J Sun, and J Zhao, "A 3d convolutional neural network framework for polyp candidates detection on the limited dataset of ct colonography," in *EMBC*. IEEE, 2018.

[13] J Tan, Y Gao, Z Liang, W Cao, M Pomeroy, Y Huo, L Li, M Barish, A Abbasi, and P Pickhardt, "3d-glcm cnn: A 3-dimensional gray-level co-occurrence matrix-based cnn model for polyp classification via ct colonography," *IEEE transactions on medical imaging*, 2019.

[14] T Uemura, J Näppi, T Hironaka, H Kim, and H Yoshida, "Comparative performance of 3d-densenet, 3d-resnet, and 3d-vgg models in polyp detection for ct colonography," in *Medical Imaging 2020: Computer-Aided Diagnosis*. SPIE, 2020.

[15] S Grosu, P Wesp, A Graser, S Maurus, C Schulz, T Knösel, C Cyran, J Ricke, M Ingrisch, and P Kazmierczak, "Machine learning–based differentiation of benign and premalignant colorectal polyps detected with ct colonography in an asymptomatic screening population: a proof-of-concept study," *Radiology*, 2021.

[16] Philipp Wesp, Sergio Grosu, Anno Graser, Stefan Maurus, Christian Schulz, Thomas Knösel, Matthias P Fabritius, Balthasar Schachtner, Benjamin M Yeh, Clemens C Cyran, et al., "Deep learning in ct colonography: differentiating premalignant from benign colorectal polyps," *European Radiology*, vol. 32, no. 7, pp. 4749–4759, 2022.

[17] S Zhang, J Wu, E Shi, S Yu, Y Gao, L Li, L Kuo, M Pomeroy, and Z Liang, "Mm-glcm-cnn: A multi-scale and multi-level based glcm-cnn for polyp classification," *CMIG*, 2023.

[18] K Smith, K Clark, W Bennett, T Nolan, J Kirby, M Wolfsberger, J Moulton, B Vendt, and J Freymann, "Data from ct colonography. the cancer imaging archive (2015)," .

[19] HIROMI Shinya and WILLIAM I Wolff, "Morphology, anatomic distribution and cancer potential of colonic polyps.," *Annals of surgery*, vol. 190, no. 6, pp. 679, 1979.

[20] Zina Ravindran, Nisha S Das, et al., "Automatic segmentation of colon using multilevel morphology and thesholding," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2021, pp. 1–4.

[21] J Nappi, A Dachman, P MacEneaney, and H Yoshida, "Effect of knowledge-guided colon segmentation in automated detection of polyps in ct colonography," in *Medical Imaging: Physiology and Function from Multidimensional Images*, 2002.

[22] Asem M. Ali, Aly A. Farag, and Naif A. Alajlan, "Multimodal imaging: modelling and segmentation with biomedical applications," *IET Computer Vision*, vol. 6, pp. 524–539, 2012.

[23] A Bochkovskiy, C Wang, and H Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv 2004.10934*, 2020.

[24] Mohammadreza Zolfaghari, Gabriel L Oliveira, Nima Sedaghat, and Thomas Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2904–2913.

[25] Haiying Liu, Fengqian Sun, Jason Gu, and Lixia Deng, "Sf-yolov5: A lightweight small object detection algorithm based on improved feature fusion mode," *Sensors*, vol. 22, no. 15, pp. 5817, 2022.

[26] C Wang, A Bochkovskiy, and H Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv 2207.02696*, 2022.

[27] S Ren, K He, R Girshick, and J Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NIPS*, 2015.

[28] H Zhang, H Chang, B Ma, N Wang, and X Chen, "Dynamic r-cnn: Towards high quality object detection via dynamic training," in *ECCV*, 2020.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[30] P Sun, R Zhang, Y Jiang, T Kong, C Xu, W Zhan, M Tomizuka, L Li, Z Yuan, C Wang, et al., "Sparse r-cnn: End-to-end object detection with learnable proposals," in *CVPR*, 2021.

[31] Z Liu, Y Lin, Y Cao, H Hu, Y Wei, Z Zhang, S Lin, and B Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[32] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[33] K He, X Zhang, S Ren, and J Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[34] K Simonyan and A Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv 1409.1556*, 2014.

[35] C Szegedy, W Liu, Y Jia, P Sermanet, S Reed, D Anguelov, D Erhan, V Vanhoucke, and A Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[36] G Huang, Z Liu, L Van Der Maaten, and K Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[37] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[38] C Liu, B Zoph, M Neumann, J Shlens, W Hua, L Li, L Fei-Fei, A Yuille, J Huang, and K Murphy, "Progressive neural architecture search," in *ECCV*, 2018.

[39] M Sandler, A Howard, M Zhu, A Zhmoginov, and L Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[40] S Xie, R Girshick, P Dollár, Z Tu, and K He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[41] N Ma, X Zhang, H Zheng, and J Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, 2018.