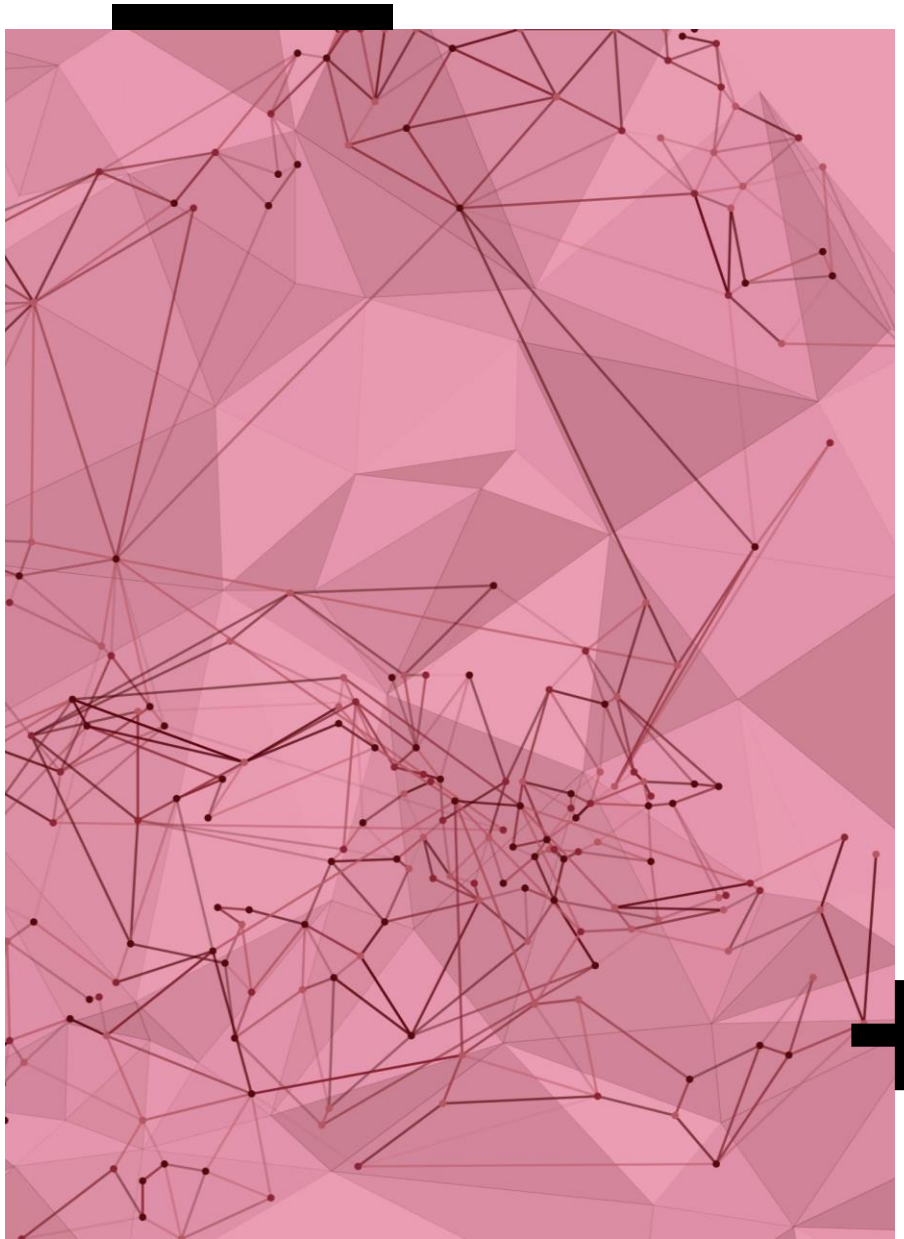


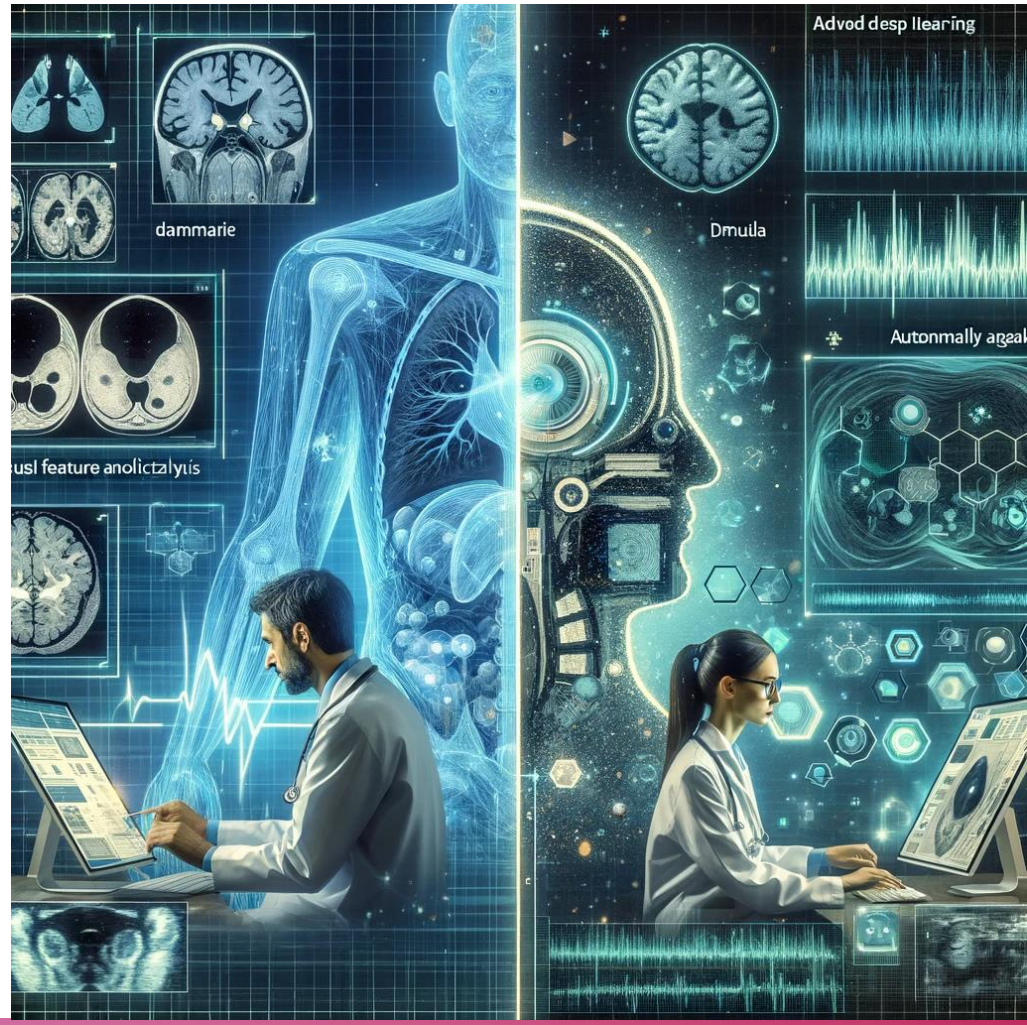
Evaluation of Multimodal Models: Assessing Performance and Finding Improvements

Metamorphic Testing for Medical Image Analysis

WU, Haoran WU, Yushan

Multimodality and Healthcare

- Multimodal: integration of multiple modes of communication and interaction.
- Application fields: healthcare, education, finance etc.
- Further improve the use of AI in healthcare.



AI in Medical Imaging

- Medical errors are a critical issue.
- A leading cause is diagnostic errors.
- AI can enhance the accuracy of medical diagnosis tools.
- AI-enabled tools monitor vital signs and calculate early warning scores to identify signs of events.

Challenges in AI-Driven Medical Diagnosis

- Fallibility: AI misalignment with clinicians' assessments.
- Need for reliable and robust testing frameworks.
- The methodologies for generating test cases in general computer vision software cannot be directly applied due to the complex nature of medical diagnosis.

Introducing MedTest

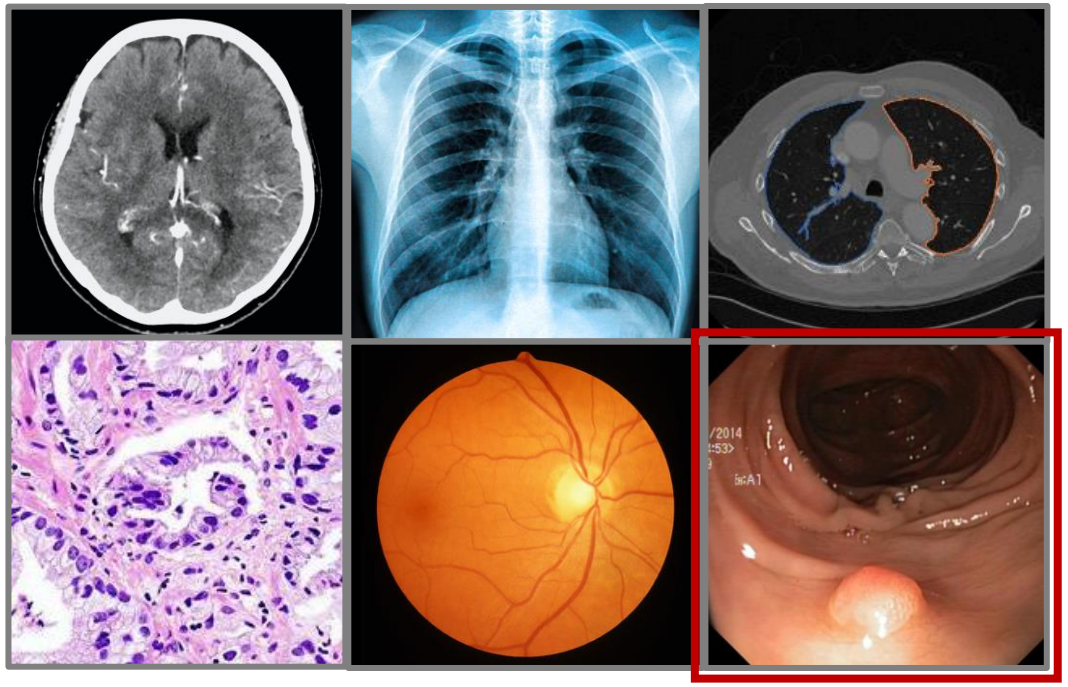
- **MedTest**: A novel metamorphic testing paradigm targeting models on medical imaging tasks.

- Conducted a pilot study, revealing 9 metamorphic relations, across four artifact categories: lightness, motion, object artifacts, and non-object artifacts.

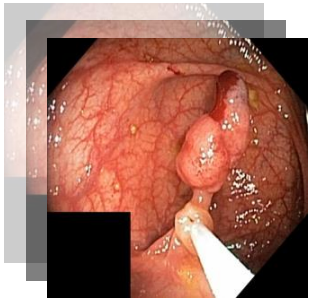
- Testing in both commercial software and state-of-the-art algorithms.

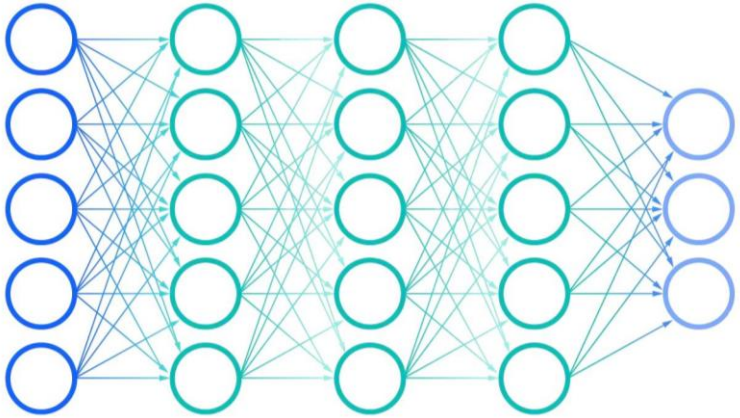
Medical Image Analysis

- Various imaging techniques to understand medical data.
- Computer-assistance methods enhance diagnosis accuracy and efficiency.
- Increased application of deep learning methods.

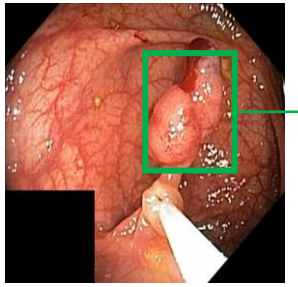


Common Tasks



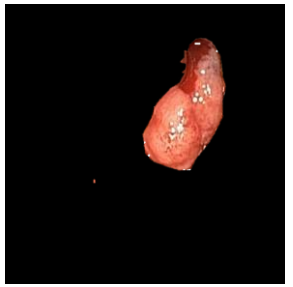


Output

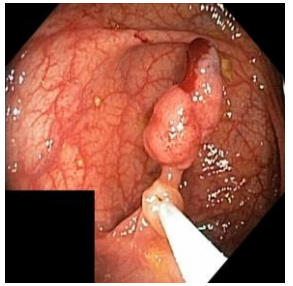


Polyp detected.
[p%]

Detection



Segmentation



Anomalies Class?

Yes + Polyp.

Classification

Metamorphic Testing

- Key idea: Automatically generate test cases to solve the test oracle problem via Metamorphic Relations (MR).
- MRs delineate the expected relationship between different sets of input-output pairs of a software application.
- Let p be a representation mapping program inputs into program outputs, and f_I and f_O are two functions for transforming the input and output domain, respectively.
- MR formulation:

$$\forall\, i, p[\![f_I(i)]\!] = f_O(p[\![i]\!])$$

Metamorphic Testing on AI models

- In our testing scenarios, let $Model$ be the model or software we target, that continuously maps each image into predicted output (e.g. segmentation mask).
- Given the original image stream $\mathbb{I}$, we can define various image perturbations $\mathbb{P}$ that simply add some artifacts and do not impact the clinical diagnosis for each image $i \in \mathbb{I}$.
- In this way, we use the following MR to test the models with additional perturbations:

$$\forall\, i \in \mathbb{I} \wedge \forall\, p \in \mathbb{P}, Model[\![p(i)]\!] \approx Model[\![i]\!]$$
$$|Model[\![p(i)]\!] - Model[\![i]\!]| < \varepsilon$$

where ε denotes a certain degree of error-tolerant rate.

Perturbation Types

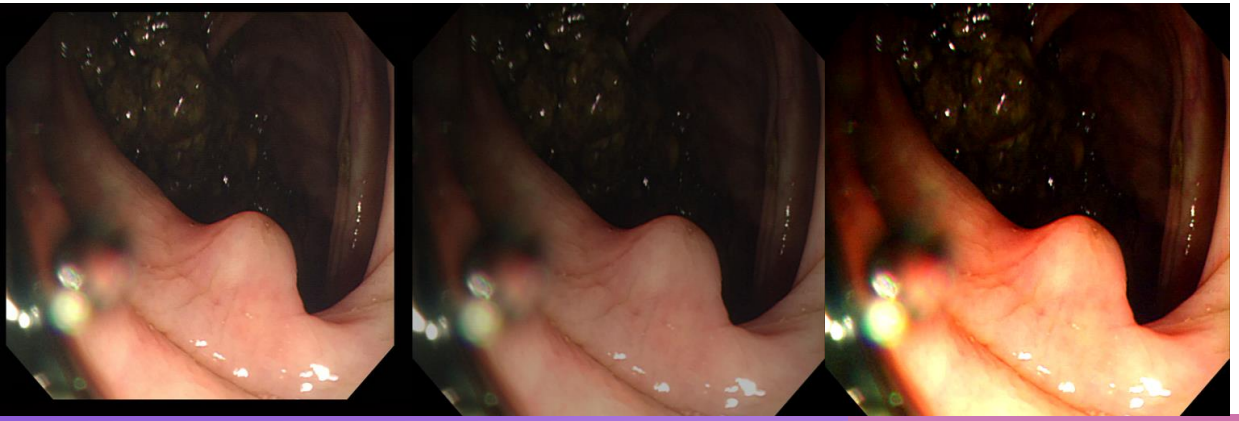
- Goal: The “seed” image and “perturbed” counterparts should yield consistent prediction results (e.g. classification label, segmentation masks).
- Perturbation criteria: clinical-semantic-preserving, realistic, unambiguous.

Perturbation Group	Type	Description
Lighting	Saturation	Over-saturation caused by excessive lighting
	Contrast	Resulting from underexposure or obstructions in the field of view
	White Balance	Color distortions due to presence of white objects
	Specularity	Reflections resembling a mirror-like surface
Motion	Blur	Blurring from hand movements or rapid camera motion
Objects	Instrument	Presence of surgical instruments in the image frame
	Feces	Incomplete colon cleansing in patients
	Blood	Visible bleeding from wounds
Non-objects	Text	Embedded clinical information related to patients

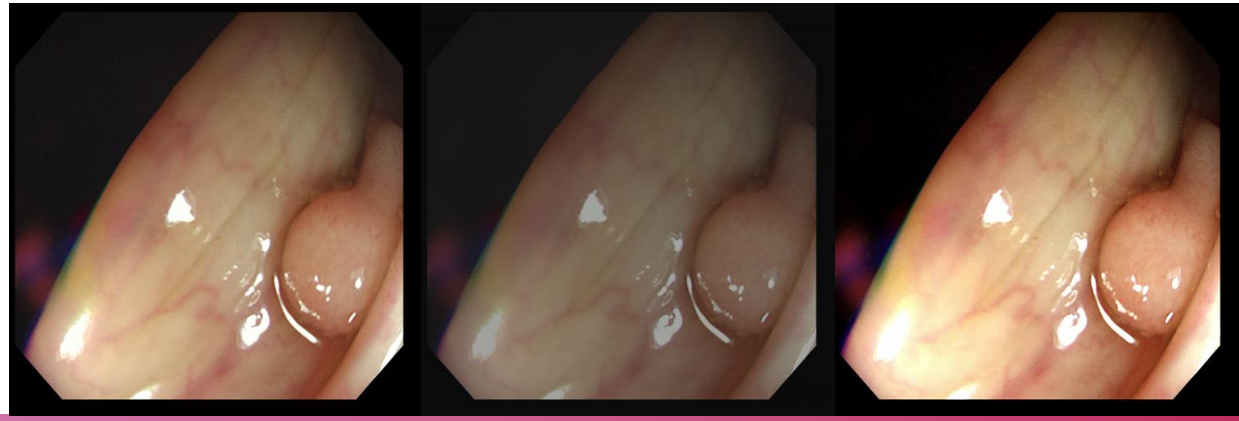
Contrast/Saturation

- The light source is too far/close to the tissue.
- Applied torchvision.transforms to adjust contrast/saturation with a random factor.

Original Contrast Saturation



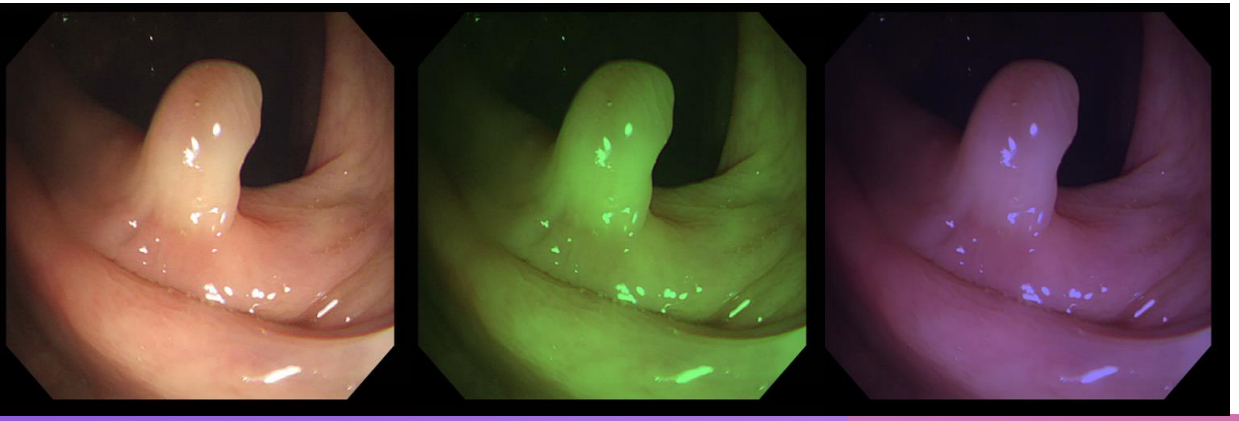
Original Contrast Saturation



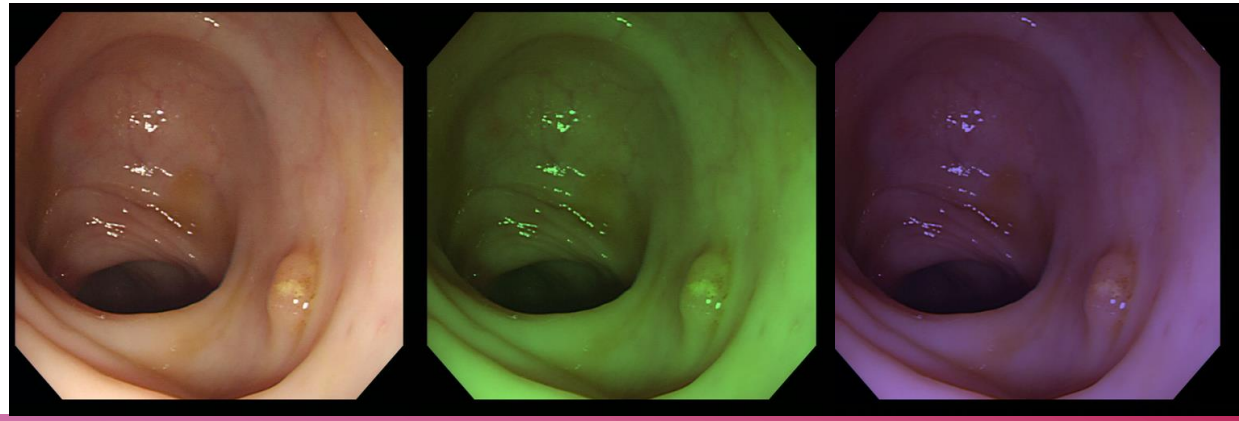
White Balance

- The white balance settings of the endoscopic camera or the lighting conditions within the endoscopic environment.
- Selectively modified the RGB channels.

Original Green Purple



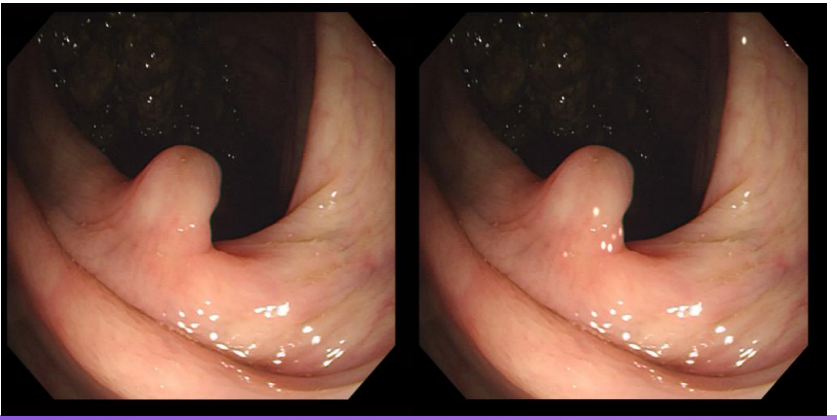
Original Green Purple



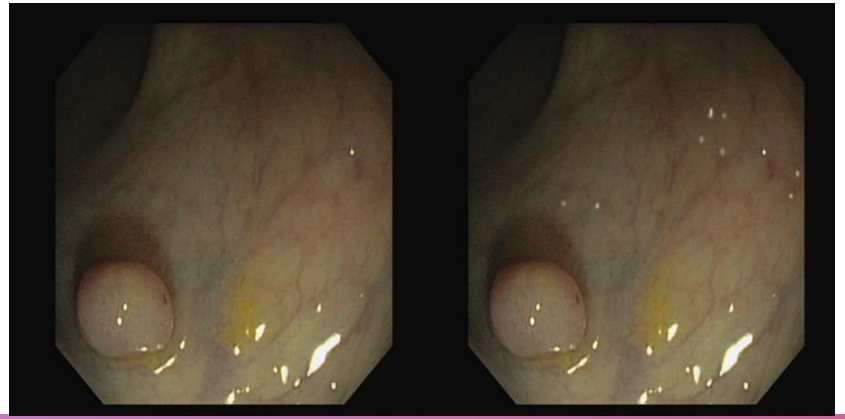
Specularity

- Resembles the specular reflection.
- Identifying clusters as potential sites, generating ellipses near the cluster centers.
- Integrated these spots with a gray mask and application of Gaussian blur.

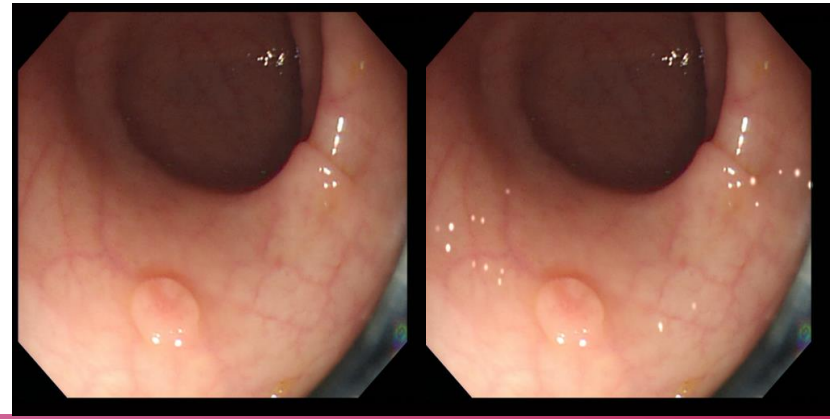
Original Specularity



Original Specularity



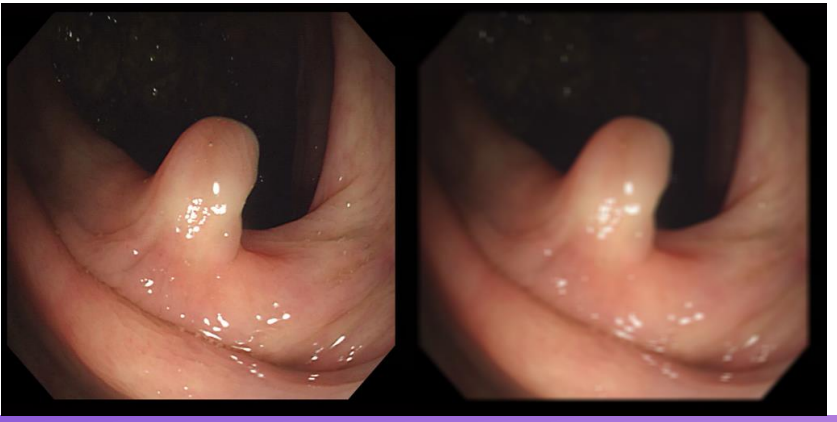
Original Specularity



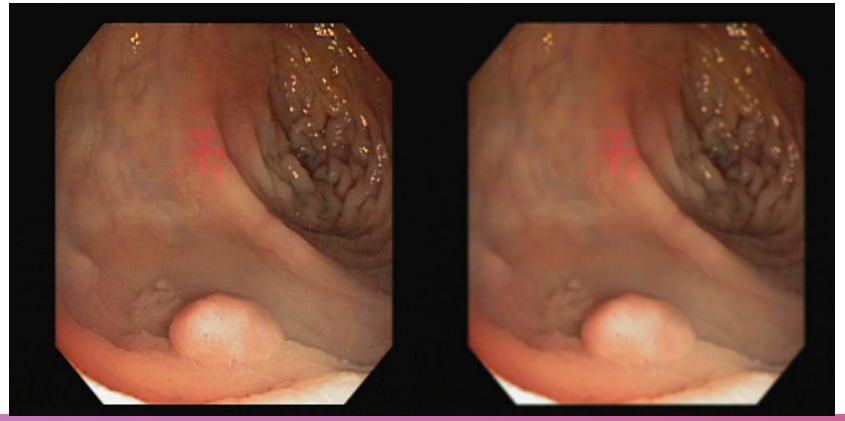
Motion Blur

- Camera movement and tissue movement.
- Employed Gaussian blur with a random factor.

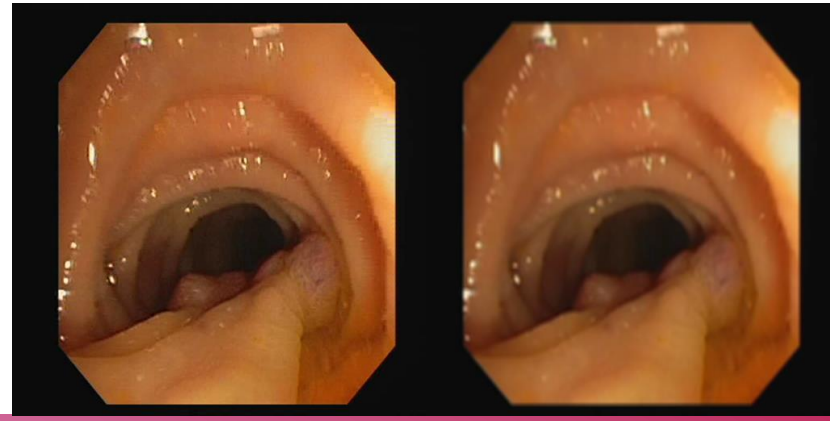
Original Blur



Original Blur



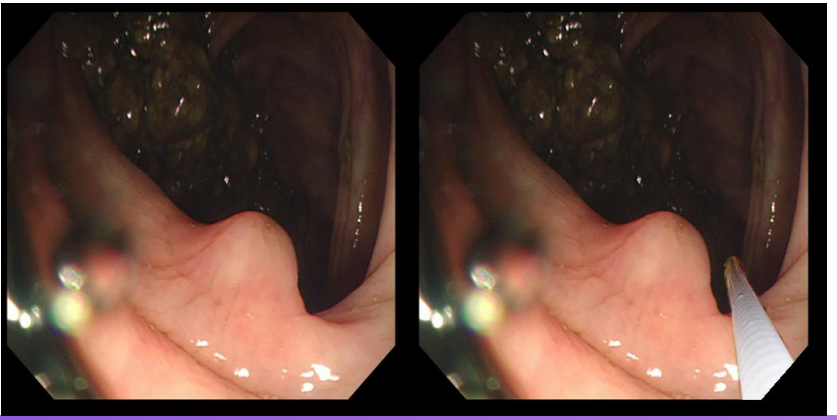
Original Blur



Instrument

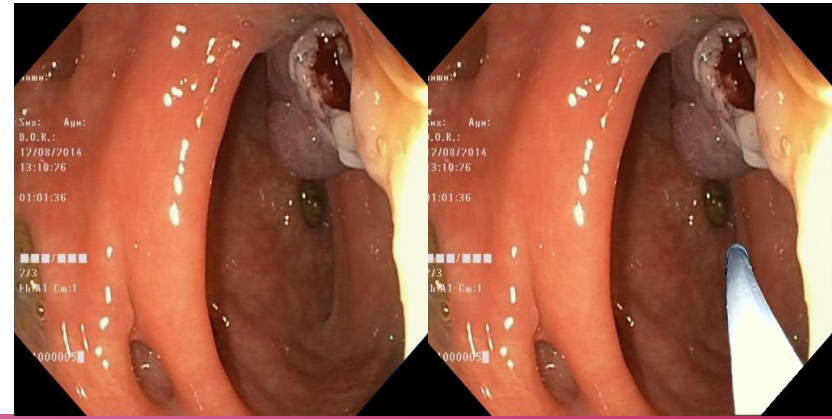
- Resembles the medical instruments that appear in operations.
- Segmented the instrument from the Kvasir-Instrument dataset.
- Utilized our algorithm to select the proper location and orientation and blend the edge.

Original Instrument



Original Instrument

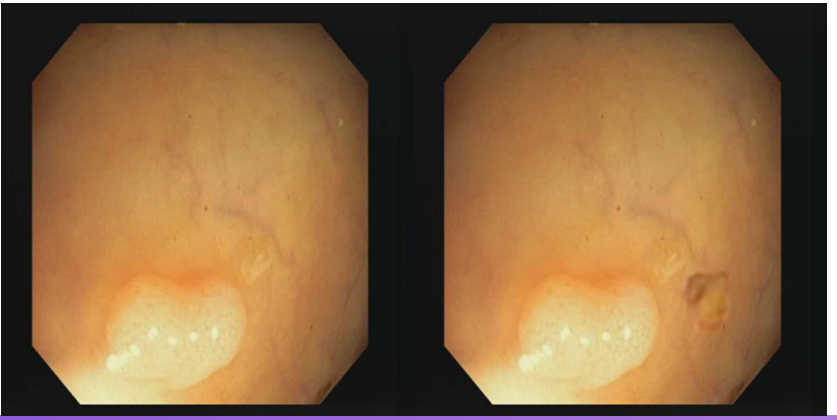
Original Instrument



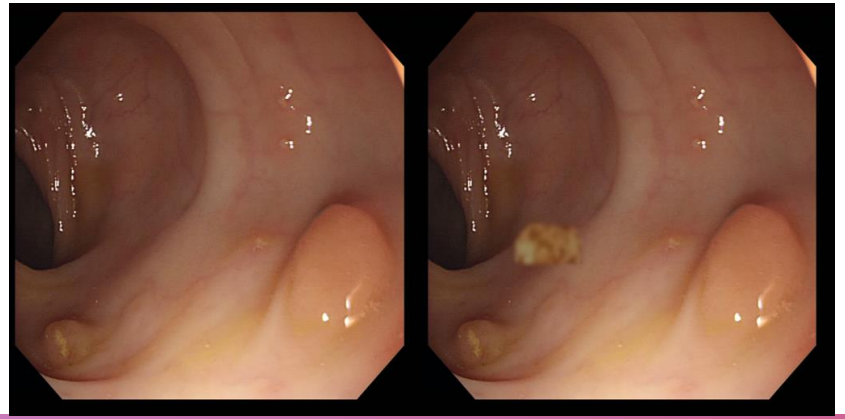
Feces

- Fecal matter appears in operations.
- Segmented with Meta's Segment Anything from Kvasir dataset.
- Utilized our algorithm to select proper location and calculated size and brightness factor to blend in.

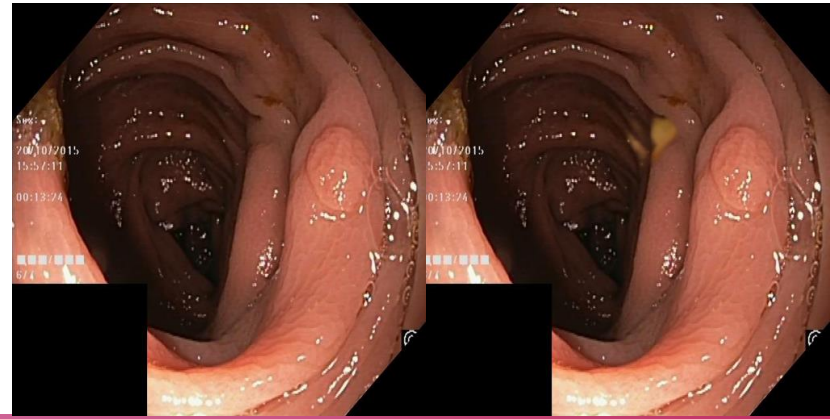
Original Feces



Original Feces



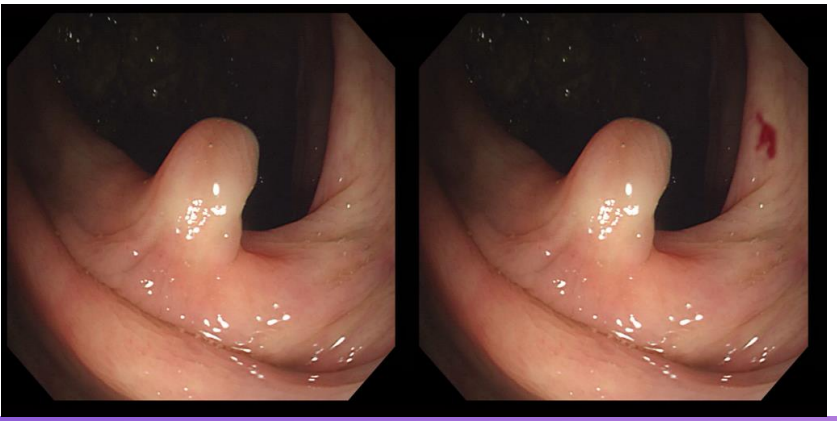
Original Feces



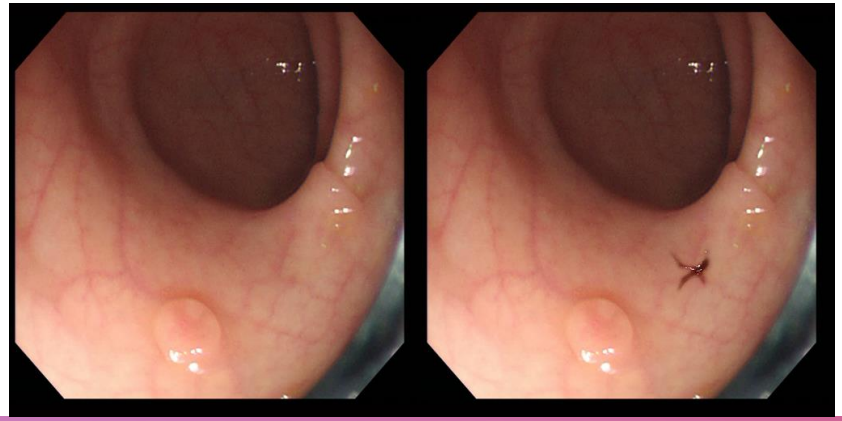
Blood

- Tissue bleeding in operations.
- Segmented the blood from EAD2020 dataset.
- Utilized our algorithm to select proper location and calculated size and brightness factor to blend in.

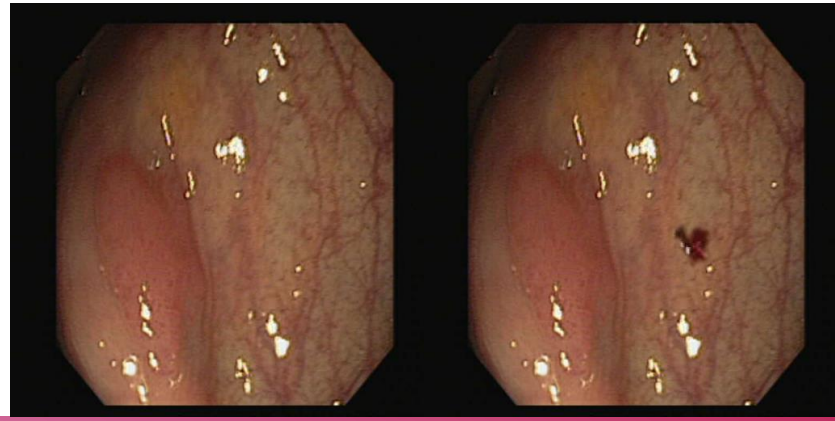
Original Blood



Original Blood



Original Blood



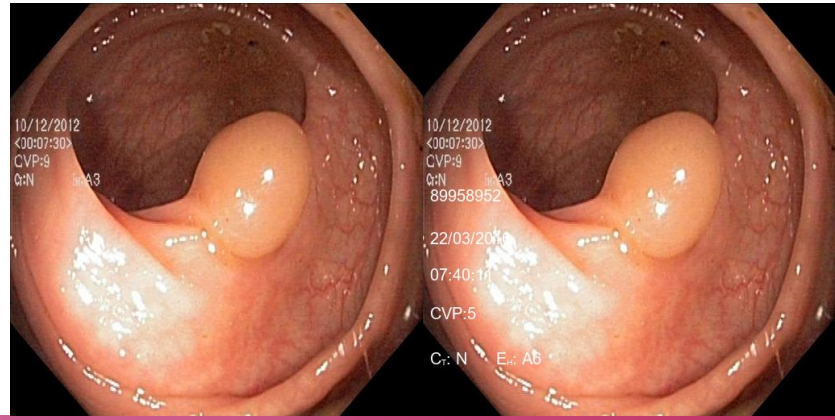
Text

- Pattern in the text displayed on endoscopic images.
- Used ImageDraw method of PIL to generate text.

Original Text

Original Text

Original Text



Evaluation

Evaluate our methodology by answering the following Research Questions (RQ):

- RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?
- RQ3: What are the various factors that influence the performance of our method and how do they do so?

Experiment Settings

- Mainly utilize clinical endoscopy images and evaluate models on polyp-related tasks.

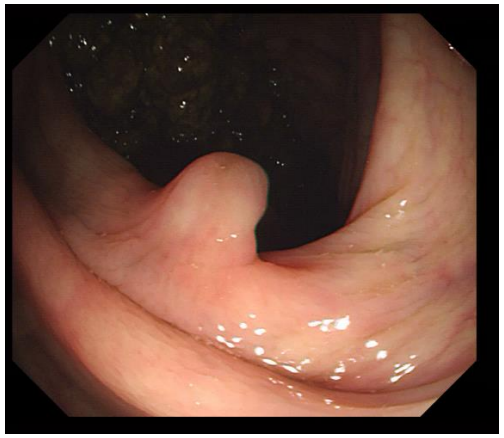
- Datasets:
  - CVC-300 (60 images), CVC-ClinicDB (612 images), CVC-ColonDB (380 images), and Kvasir (1000 images) mainly for segmentation task. 2052 seed images in total.
  - Additional Kvasir-instrument (590 images) for VQA testing.

- Models under testing:
  - Polyp **segmentation** models: PraNet, SANet, TGANet, SSFormer.
  - Multi-modal models for **Visual Question-Answering** (VQA): ChatGPT-4V.

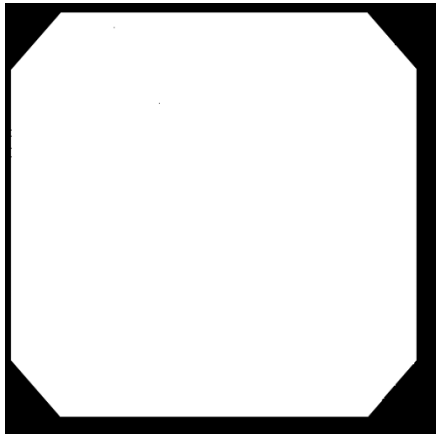
Dataset Pre-processing

- Large difference in image sizes -> Resize into 512×512.
- Extract the black frame of images to avoid possible synthesis on the edge.
- Generate gray masks for images to adjust the brightness condition of synthesized parts.

Original



Crop Mask



Gray Mask

Evaluation-RQ1

Evaluate our methodology by answering the following Research Questions (RQ):

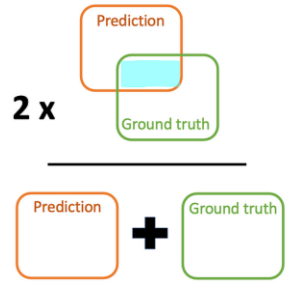
- **RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?**
- RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?
- RQ3: What are the various factors that influence the performance of our method and how do they do so?

Evaluation Criteria

Measurement for segmentation task:

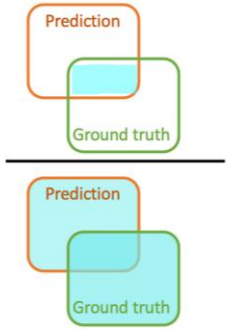
- Dice Score:

$$Dice(\hat{Y}, Y) = \frac{2 \times |\hat{Y} \cap Y|}{|\hat{Y}| + |Y|} = \frac{2 \times TP}{(TP + FP) + (TP + FN)} = \frac{2\,X\,Area\ of\ overlap}{Total\ area} =$$



- Intersection over Union (IoU) Score:

$$IoU(\hat{Y}, Y) = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|} = \frac{TP}{TP + FP + FN} = \frac{Area\ of\ overlap}{Area\ of\ union} =$$



Evaluation Criteria

Measurement for “Misclassified”/ “Error”:

- The difference between model’s performance on “seed” image and on perturbations should not exceed an error-tolerant threshold t.
- Performance is calculated by Dice/IoU Score.
- The sample counts toward an error if

$$\frac{Original\ Score\ -Artifact\ Score}{Original\ Score} > t$$

Error Finding Rate (EFR):

$$EFR = \frac{\#\ of\ error\ test\ cases}{\#\ of\ generated\ test\ cases} \times\ 100\%$$

Results

- For illustration, we choose $t = 0.25$.
- The EFRs are organized by each model, together with separate values for each dataset and perturbations.

PraNet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	3.3	6.7	0.5	1.0	4.0	5.0	0.5	0.6
Feces	0.0	1.7	0.8	2.0	7.4	9.2	0.5	1.5
Instrument	6.7	11.7	4.1	5.6	12.1	14.0	0.4	1.1
Spot	1.7	1.7	0.5	0.5	3.2	4.2	0.1	0.5
Saturation	**8.3**	**13.3**	1.6	3.4	6.6	8.4	5.8	9.6
Contrast	1.7	5.0	0.3	0.8	4.7	6.1	1.3	2.2
White Balance	**8.3**	**13.3**	**12.7**	**18.0**	**19.8**	**22.7**	7.5	12.3
Blur	**8.3**	8.3	9.6	13.6	14.2	17.2	**14.2**	**18.8**
Text	0.0	0.0	0.7	0.8	5.0	5.8	0.2	0.3

PraNet: Overall EFR = 4.38%

SANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	0.0	0.0	0.0	0.0	2.9	3.4	0.1	0.1
Feces	1.7	1.7	0.3	0.5	6.9	7.4	0.1	0.2
Instrument	1.7	1.7	0.2	0.7	5.5	5.8	0.0	0.0
Spot	0.0	0.0	0.2	0.3	4.2	4.5	0.0	0.0
Saturation	**5.0**	**6.7**	1.0	1.8	3.4	5.5	1.7	3.1
Contrast	0.0	0.0	0.2	0.2	3.4	4.0	0.0	0.2
White Balance	0.0	0.0	**3.4**	**6.2**	**10.8**	**14.0**	**5.4**	**9.2**
Blur	3.3	5.0	0.3	0.5	6.3	8.7	1.1	2.0
Text	0.0	0.0	0.5	1.0	5.5	5.8	0.0	0.1

SANet: Overall EFR = 1.70%

Results

SSFormer	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	3.3	3.3	0.2	0.2	5.0	5.3	0.1	0.1
Feces	0.0	0.0	0.3	0.5	7.6	8.2	0.0	0.0
Instrument	3.3	6.7	1.8	2.5	7.1	7.6	0.0	0.0
Spot	0.0	0.0	0.3	0.3	2.4	2.4	0.0	0.0
Saturation	**6.7**	**10.0**	1.0	1.3	2.6	4.5	0.4	0.8
Contrast	1.7	3.3	0.2	0.2	3.9	4.7	0.3	0.5
White Balance	3.3	5.0	**4.7**	**7.5**	**11.8**	**13.9**	**2.0**	**4.0**
Blur	0.0	1.7	0.2	0.2	3.4	3.4	0.3	0.4
Text	0.0	0.0	0.3	0.3	2.1	2.6	0.0	0.1

SSFormer: Overall EFR = 1.47%

TGANet	CVC-300		CVC-ClinicDB		CVC-ColonDB		Kvasir	
t=0.25	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
Blood	16.7	20.0	15.8	22.1	23.9	29.2	12.9	15.7
Feces	13.3	25.0	4.4	7.0	13.9	18.2	2.7	3.7
Instrument	30.0	**46.7**	9.2	14.9	18.9	24.2	4.4	6.9
Spot	3.3	3.3	1.5	2.1	5.5	6.6	0.8	1.0
Saturation	16.7	18.3	21.2	28.9	21.8	24.7	**46.1**	**53.7**
Contrast	0.0	1.7	12.9	17.3	26.8	29.2	14.3	18.0
White Balance	**31.7**	38.3	**47.5**	**59.5**	**35.3**	**40.8**	43.0	49.8
Blur	28.3	31.7	4.7	6.5	9.7	11.8	15.3	18.3
Text	8.3	8.3	3.9	5.1	10.8	13.4	3.9	5.6

TGANet: Overall EFR = 15.70%

Analysis-Perturbations

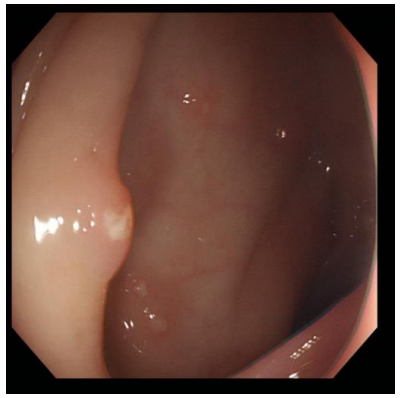
- Lighting conditions (e.g. white balance, saturation) can trigger most errors.
  - Possible explanation: 1. Edges of polyps become vague. 2. Sensitive to color.

- Motion Blurring can also lead to some corner cases.
  - Possible explanation: Edges of polyps become vague.

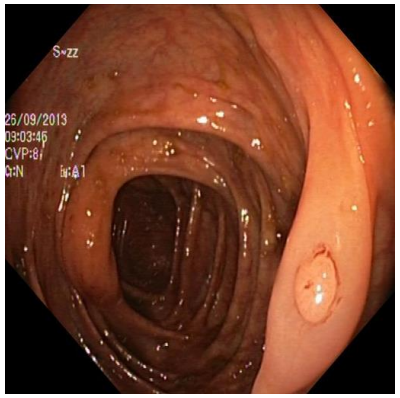
- Instrument perturbation resulted in some misleading cases.
  - Possible explanation: Unseen elements in the training data.

Analysis-Datasets & Models

- Datasets: Perturbations generated on CVC-ColonDB and Kvasir led to higher EFR.
  - CVC-ColonDB is relatively new and not often used in training.
  - Kvasir has different image layouts compared to CVC datasets.



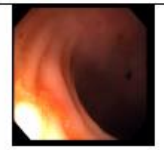
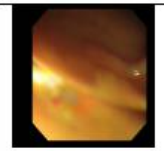
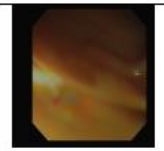
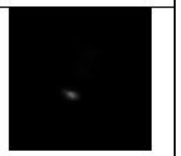
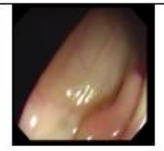
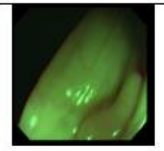
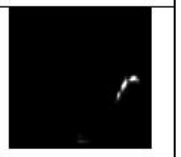
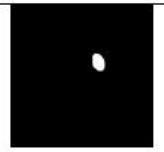
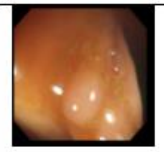
CVC-ColonDB sample

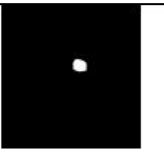
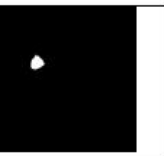
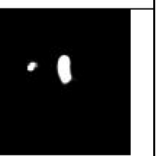
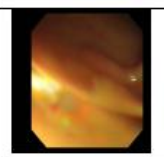
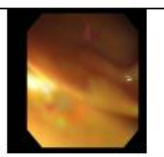
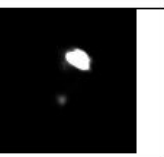
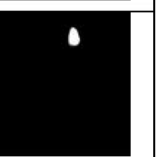


Kvasir sample

- Models: Highest overall EFR generated from TGANet.
  - Possible Explanation: TGANet incorporates additional auxiliary classification tasks for polyp descriptions, which may lead to overfitting.

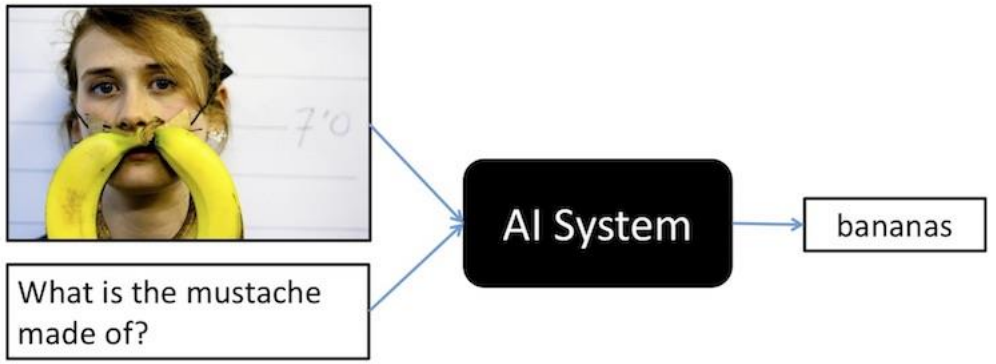
Visualization

Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Saturation					
Contrast					
White-Balance					
Specularity					
Blur					

Artifact	Original Image	Image with Artifact	Ground Truth	Output (Original)	Output (Artifact)
Instrument					
Feces					
Blood					
Text					

Visual Question-Answering (VQA)

- VQA refers to the task of answering open-ended questions based on an image.
- These questions require an understanding of vision, language, and commonsense knowledge to answer.



VQA

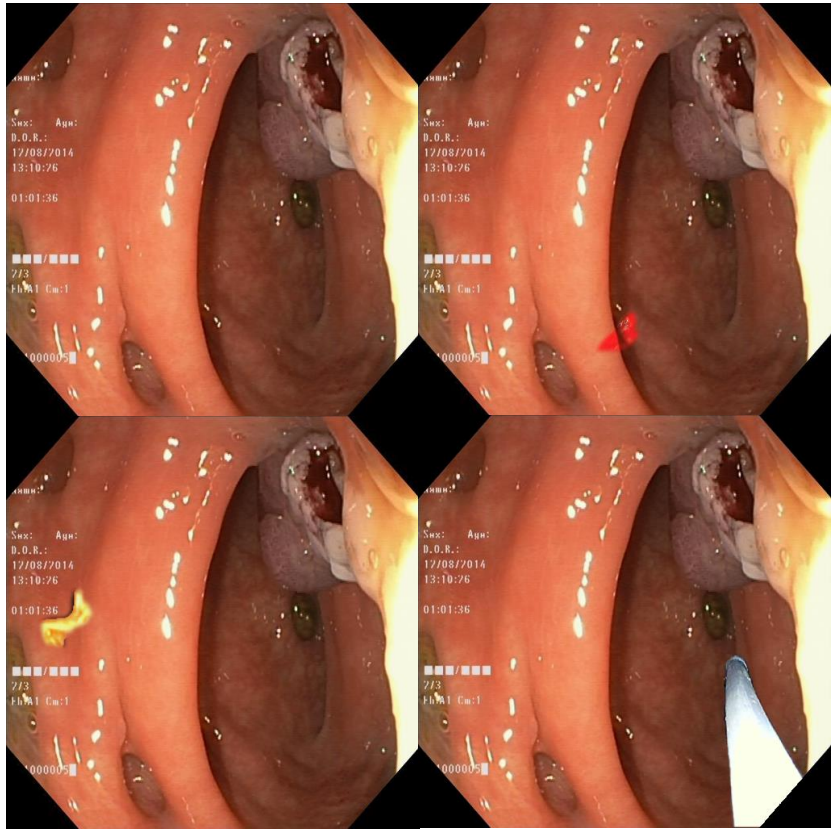
- Tested on GPT-4V
- Used the questions provided in the CLEF2023 MEDVQA Dataset.
- On going experiment process.

Question Number	Question
1	Are there any abnormalities in the image?
2	Are there any anatomical landmarks in the image?
3	Are there any instruments in the image?
4	Have all polyps been removed?
5	How many findings are present?
6	How many instruments are in the image?
7	How many polyps are in the image?
8	Is there a green/black box artefact?
9	Is there text?
10	Is this finding easy to detect?
11	What color is the abnormality?
12	What color is the anatomical landmark?
13	What is the size of the polyp?
14	What type of polyp is present?
15	What type of procedure is the image taken from?
16	Where in the image is the abnormality?
17	Where in the image is the anatomical landmark?
18	Where in the image is the instrument?

VQA

- Illustration on VQA testing case

Original



Blood

Feces

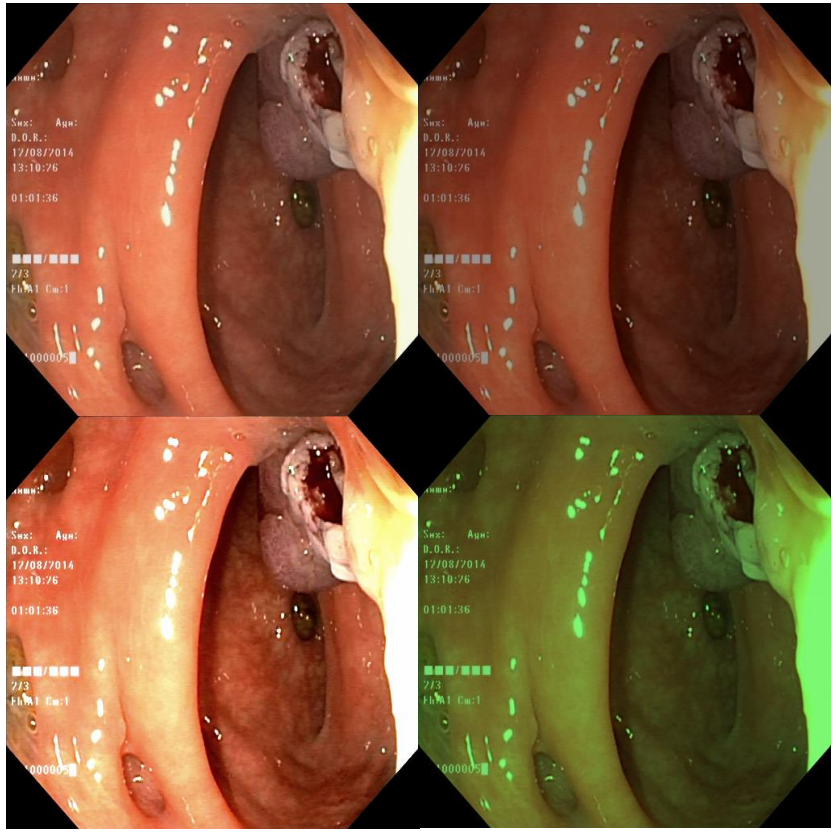
Instrument

Question	Ground Truth	Original	Blood	Feces	Instrument
Are there any abnormalities in the image?	Polyp	Polyp	Bleeding	Feces	Polyp
Are there any anatomical landmarks in the image?	No	No	No	Yes	Yes
Are there any instruments in the image?	No	No	No	No	Yes
Have all polyps been removed?	No	No	Not relevant	Not relevant	No
How many findings are present?	1	1	1	1	1
How many instruments are in the image?	0	0	0	0	1
How many polyps are in the image?	1	1	0	0	1
Is there a green/black box artefact?	No	No	No	No	No
Is there text?	Yes	Yes	Yes	Yes	Yes
Is this finding easy to detect?	Yes	Yes	Yes	Yes	Yes
What color is the abnormality?	Red, Pink, Grey	Red	Red	Brown	Red
What color is the anatomical landmark?	Not relevant	Not relevant	Not relevant	Pink	Pink
What is the size of the polyp?	>20mm	>10mm	Not relevant	Not relevant	>10mm
What type of polyp is present?	Paris is	Paris Ip	Not relevant	Not relevant	Paris Ip
What type of procedure is the image taken from?	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy
Where in the image is the abnormality?	Center, Upper-right, Center-right, Upper-center	Center-Left	Center-Left	Bottom-Center	Center-Left
Where in the image is the anatomical landmark?	Not relevant	Not relevant	Not relevant	Center	Center
Where in the image is the instrument?	Not relevant	Not relevant	Not relevant	Not relevant	Bottom-Center

VQA

- Illustration on VQA testing case

Original



Contrast

Saturation

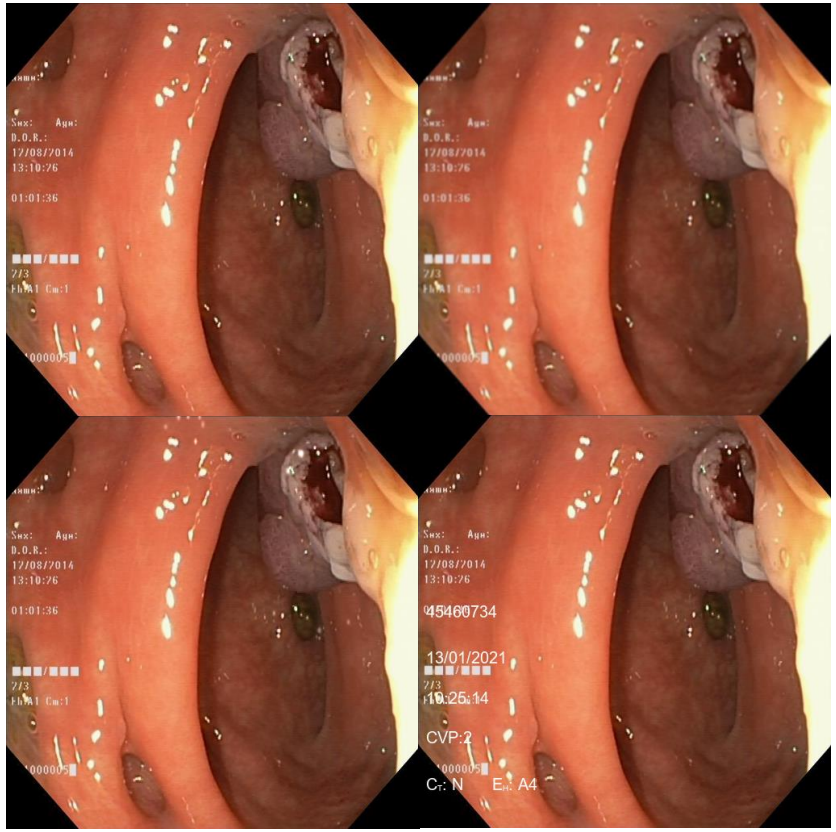
White Balance

Question	Ground Truth	Original	Contrast	Saturation	White Balance
Are there any abnormalities in the image?	Polyp	Polyp	Polyp	Polyp	Polyp
Are there any anatomical landmarks in the image?	No	No	Yes	Yes	Yes
Are there any instruments in the image?	No	No	No	No	No
Have all polyps been removed?	No	No	No	No	No
How many findings are present?	1	1	1	1	1
How many instruments are in the image?	0	0	0	0	0
How many polyps are in the image?	1	1	1	1	1
Is there a green/black box artefact?	No	No	No	No	No
Is there text?	Yes	Yes	Yes	Yes	Yes
Is this finding easy to detect?	Yes	Yes	Yes	Yes	Difficult due to color alteration
What color is the abnormality?	Red, Pink, Grey	Red	Red	Red	Not applicable due to WB
What color is the anatomical landmark?	Not relevant	Not relevant	Pink	Pink	Not applicable due to WB
What is the size of the polyp?	>20mm	>10mm	>10mm	>10mm	Not applicable due to WB
What type of polyp is present?	Paris is	Paris Ip	Paris Ip	Paris Ip	Not applicable due to WB
What type of procedure is the image taken from?	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy
Where in the image is the abnormality?	Center, Upper-right, Center-right, Upper-center	Center-Left	Center-Left	Center-Left	Not applicable due to WB
Where in the image is the anatomical landmark?	Not relevant	Not relevant	Center	Center	Not applicable due to WB
Where in the image is the instrument?	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant

VQA

- Illustration on VQA testing case

Original



Blur

Specularity

Text

Question	Ground Truth	Original	Blur	Specularity	Text
Are there any abnormalities in the image?	Polyp	Polyp	Polyp	Polyp	Polyp
Are there any anatomical landmarks in the image?	No	No	Yes	Yes	Yes
Are there any instruments in the image?	No	No	No	No	No
Have all polyps been removed?	No	No	No	No	No
How many findings are present?	1	1	1	1	1
How many instruments are in the image?	0	0	0	0	0
How many polyps are in the image?	1	1	1	1	1
Is there a green/black box artefact?	No	No	No	No	No
Is there text?	Yes	Yes	Yes	Yes	Yes
Is this finding easy to detect?	Yes	Yes	No	Yes	Yes
What color is the abnormality?	Red, Pink, Grey	Red	Red	Red	Red
What color is the anatomical landmark?	Not relevant	Not relevant	Pink	Pink	Pink
What is the size of the polyp?	>20mm	>10mm	>10mm	>10mm	>10mm
What type of polyp is present?	Paris is	Paris Ip	Paris Ip	Paris Ip	Paris Ip
What type of procedure is the image taken from?	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy	Colonoscopy
Where in the image is the abnormality?	Center, Upper-right, Center-right, Upper-center	Center-Left	Center-Left	Center-Left	Center-Left
Where in the image is the anatomical landmark?	Not relevant	Not relevant	Center	Center	Center
Where in the image is the instrument?	Not relevant	Not relevant	Not relevant	Not relevant	Not relevant

Answer-RQ1

Evaluate our methodology by answering the following Research Questions (RQ):

RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?

Answer to RQ1: MedTest obtains up to 15.70% EFR when testing on segmentation models, and qualitatively affected VQA models' performances, which indicates that MedTest can effectively discover corner cases and be used for further testing the robustness of other models.

Evaluation-RQ2

Evaluate our methodology by answering the following Research Questions (RQ):

- RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- **RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?**
- RQ3: What are the various factors that influence the performance of our method and how do they do so?

Re-training

- We plan to retrain the models with test cases synthesized by MedTest to improve the performances of those models.
- For large language models like GPT-4V, we will use prompt engineering and in context learning approach instead of training.
- The optimization of training outcomes is still ongoing.

Answer-RQ2

Evaluate our methodology by answering the following Research Questions (RQ):

RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?

Answer to RQ2: We are still in the process of re-training the academic medical image diagnosis models to achieve a better performance.

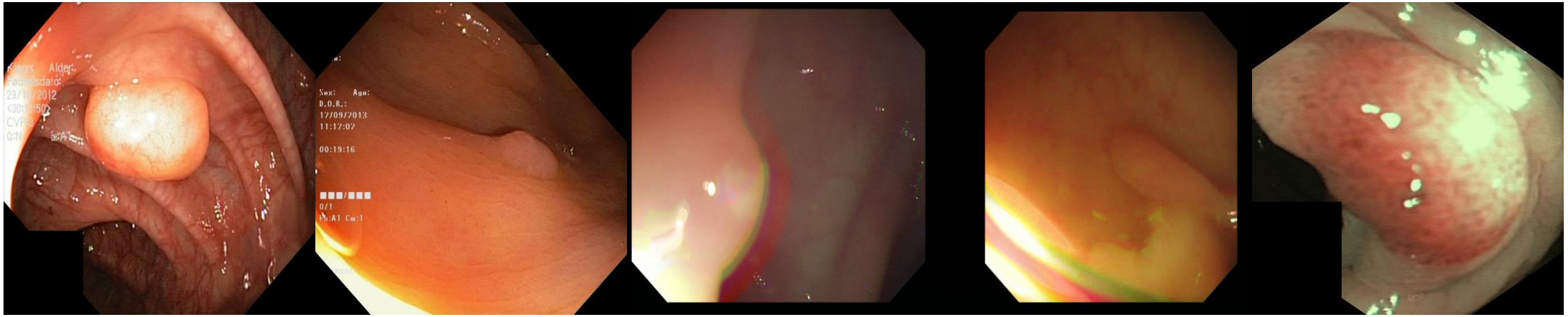
Evaluation-RQ3

Evaluate our methodology by answering the following Research Questions (RQ):

- RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?
- **RQ3: What are the various factors that influence the performance of our method and how do they do so?**

External Factors and Influences

- Divergence in Image Structure and Overlay
- Polyp Characteristics
- Lighting Conditions



Answer-RQ3

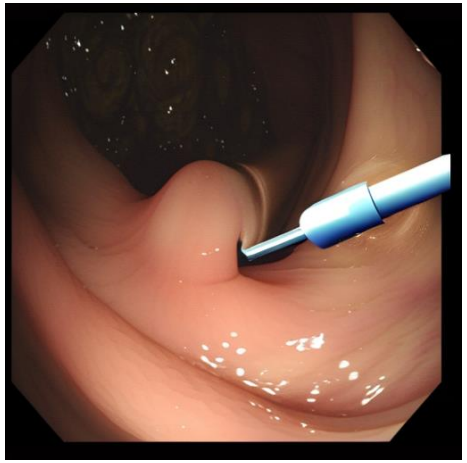
Evaluate our methodology by answering the following Research Questions (RQ):

RQ3: What are the various factors that influence the performance of our method and how do they do so?

Answer to RQ3: There are several factors related to the quality and structure of the original images that may affect the performance of MedTest.

Future Work

- Further exploration into VQA: Systematic testing on Multi-modal Large Language Models and VQA models specified on medical images.
- Retraining the models under testing for further improvements.
- Generative adversarial networks (GANs) to generate perturbations.



Conclusion

- We designed a comprehensive metamorphic testing paradigm targeting models and software on medical imaging tasks.
- With our clinical-equivalent perturbations, our method was proved to effectively identify potential model errors.
- Future work focuses on expanding the testing objectives of MedTest and identifying the potential for performance improvements on tested models.

Thank you for listening!

References

[1] Zhang, Mengshi, et al. "DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems." *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 2018.

[2] Antol, Stanislaw, et al. "Vqa: Visual question answering." *Proceedings of the IEEE international conference on computer vision*. 2015.

[3] Chen, Songqiang, Shuo Jin, and Xiaoyuan Xie. "Testing your question answering software via asking recursively." *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2021.

[4] Chen, Tsong Y., Shing C. Cheung, and Shiu Ming Yiu. "Metamorphic testing: a new approach for generating next test cases." *arXiv preprint arXiv:2002.12543* (2020).

[5] Bohr, Adam, and Kaveh Memarzadeh. "The rise of artificial intelligence in healthcare applications." *Artificial Intelligence in healthcare*. Academic Press, 2020. 25-60.

[6] Tomar, Nikhil Kumar, et al. "TGANet: Text-guided attention for improved polyp segmentation." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2022.

[7] Pei, Kexin, et al. "Deepxplore: Automated whitebox testing of deep learning systems." *proceedings of the 26th Symposium on Operating Systems Principles*. 2017.

[8] Vázquez, David, et al. "A benchmark for endoluminal scene segmentation of colonoscopy images." *Journal of healthcare engineering* 2017 (2017).

[8] Ali, Sharib, et al. "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy." *Scientific reports* 10.1 (2020): 2748.