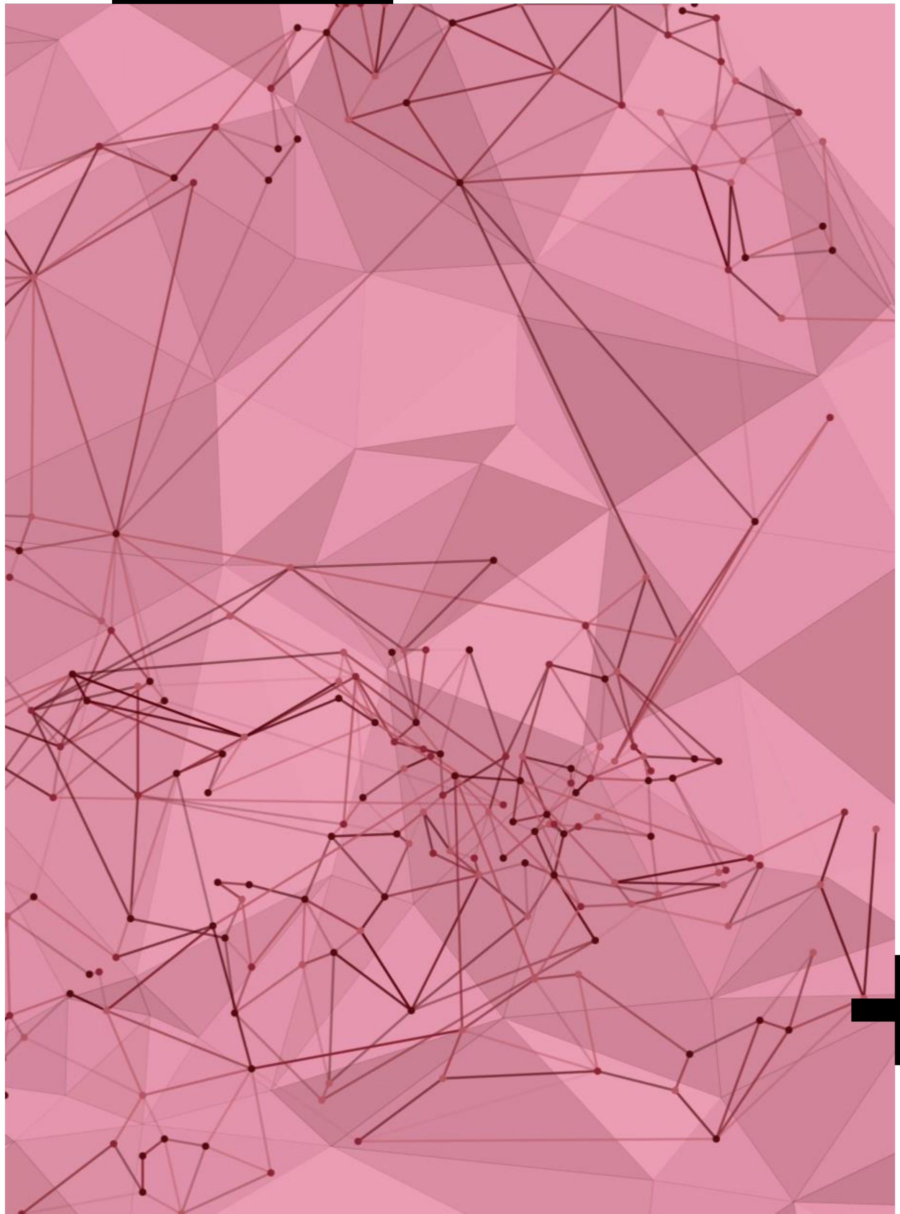


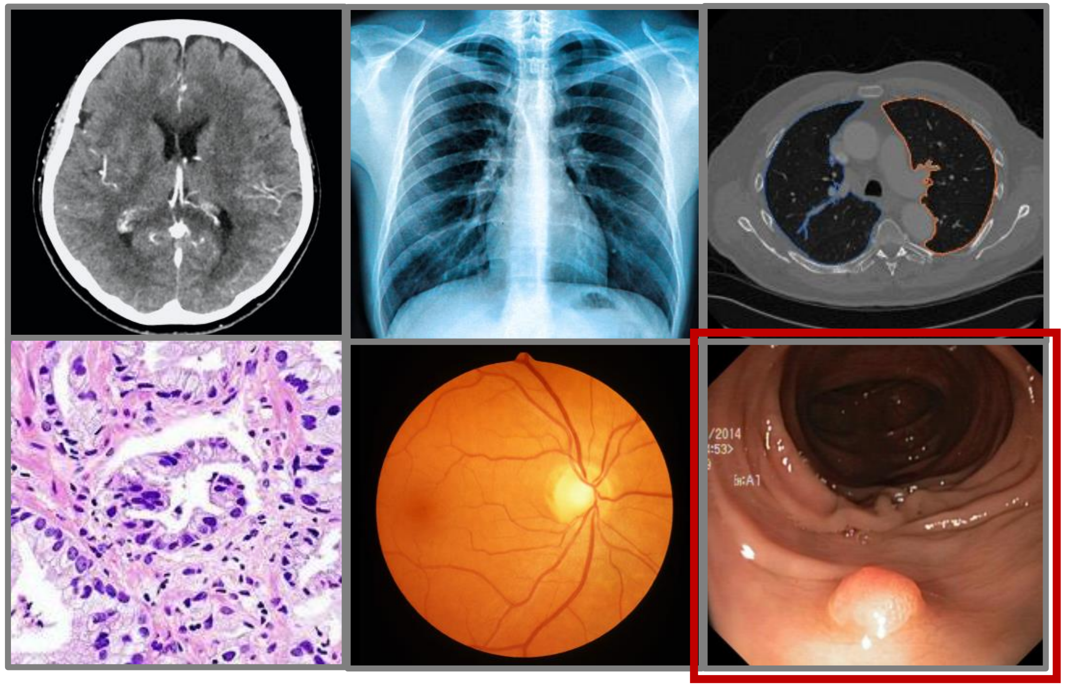
Evaluation of Multimodal Models: Assessing Performance and Finding Improvements

Metamorphic Testing for Medical Image Analysis

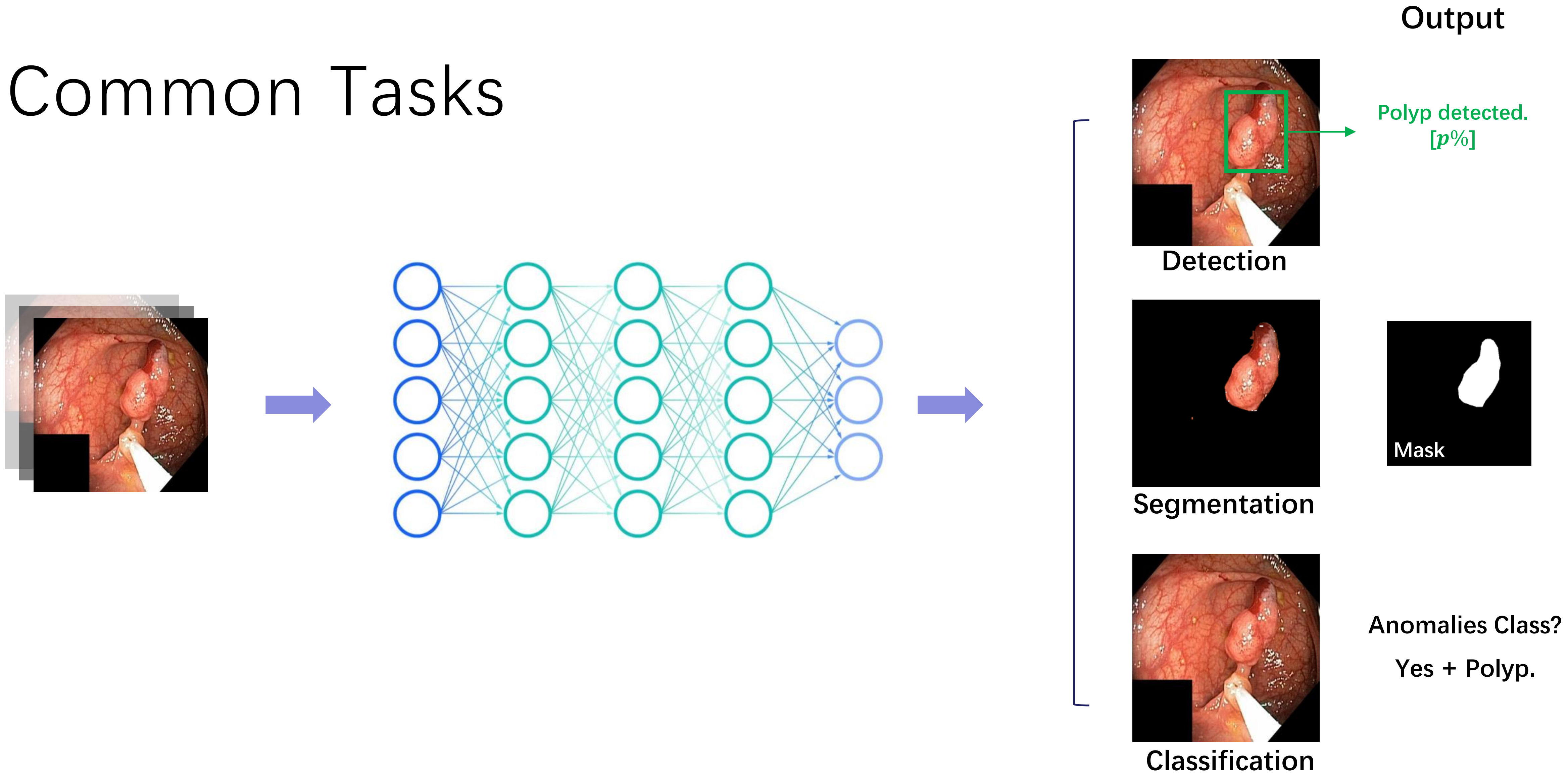
WU, Haoran WU, Yushan

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.

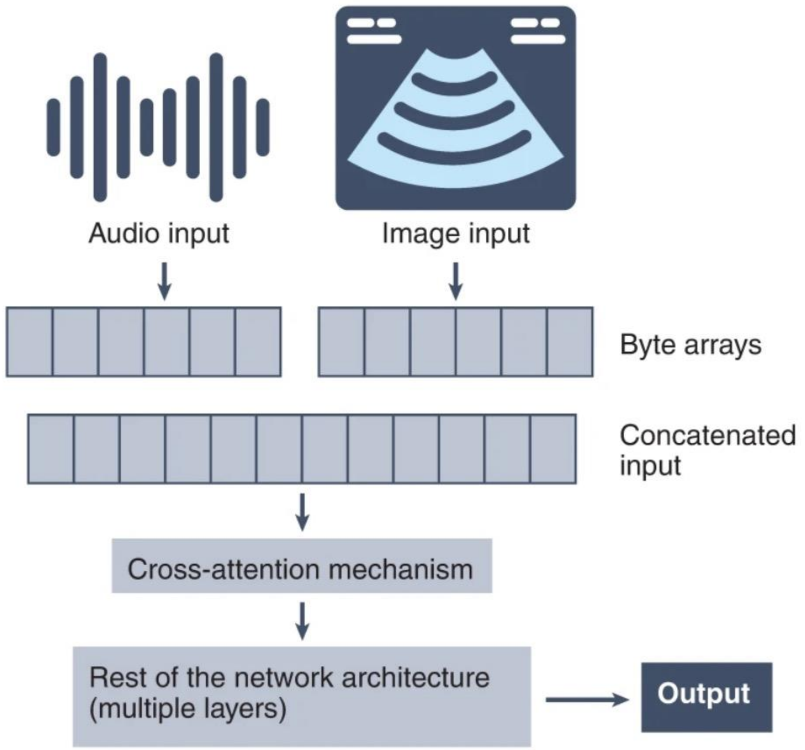


Common Tasks



Multimodality and Healthcare

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



Motivation

- 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.
- Further training on those algorithms to improve the model performance.

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:

$$\forall\, i \in \mathbb{I} \land \forall\, p \in \mathbb{P}, Model[\![p(i)]\!] \approx Model[\![i]\!]$$

$$|Model[\![p(i)]\!] - Model[\![i]\!]| < \varepsilon$$

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

Perturbation Types

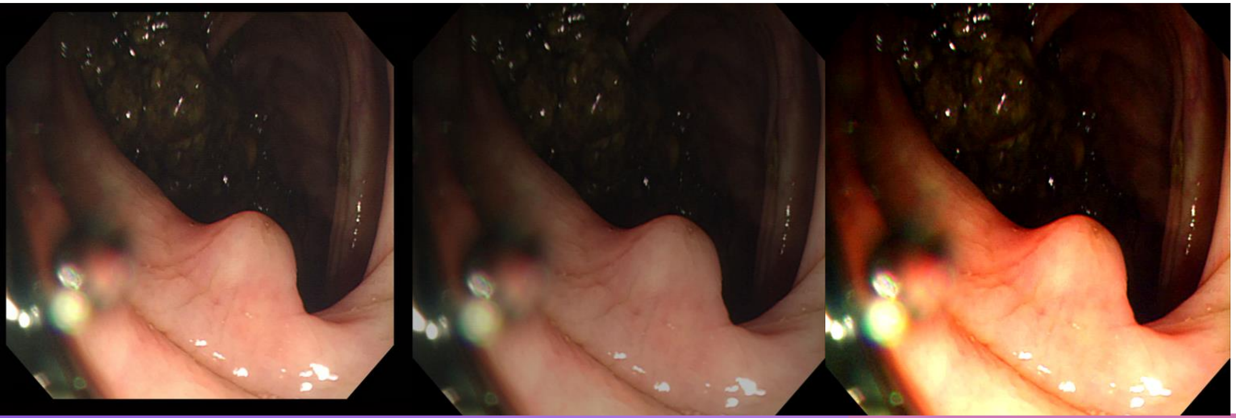
- Goal: The “seed” image and “perturbed” counterparts should yield consistent prediction results (e.g. classification label, segmentation masks).
- Perturbation criteria: clinical-semantic-preserving, realistic, unambiguous.

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

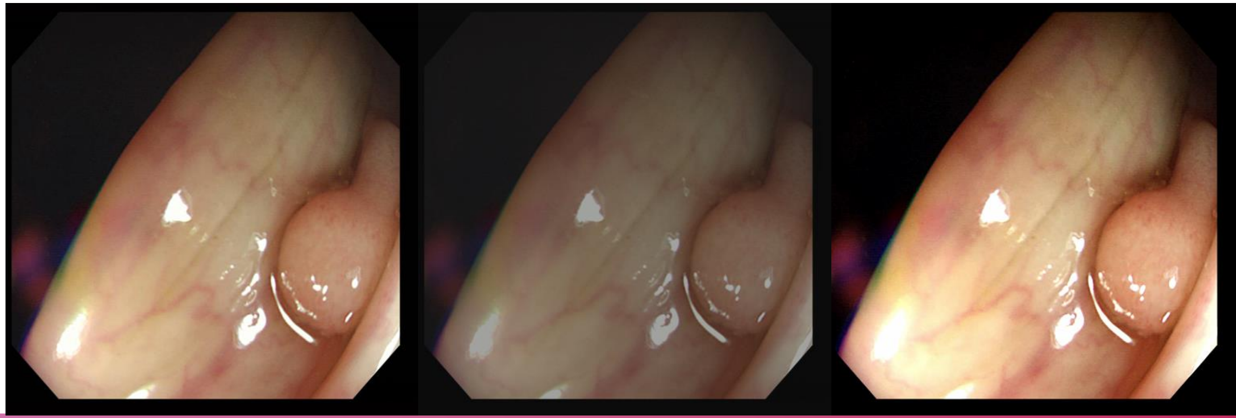
Contrast/Saturation

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

Original Contrast Saturation



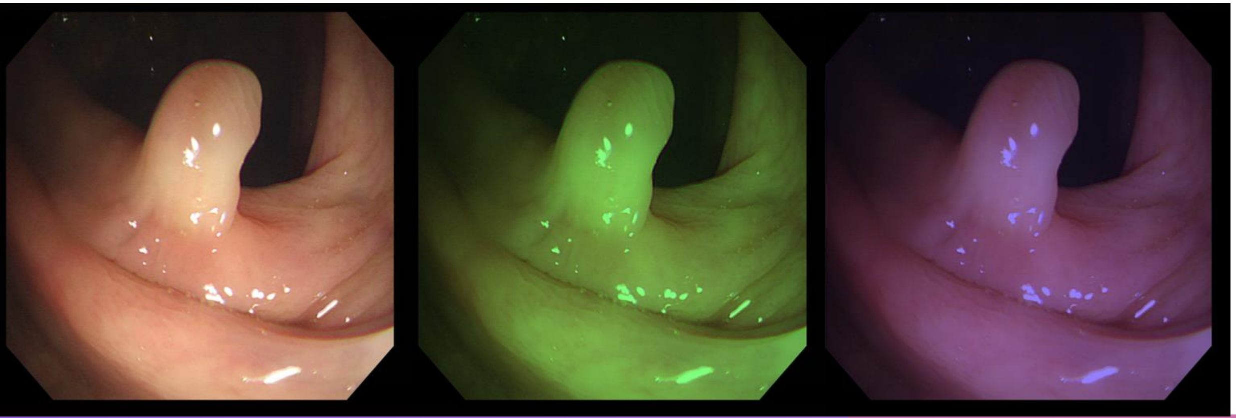
Original Contrast Saturation



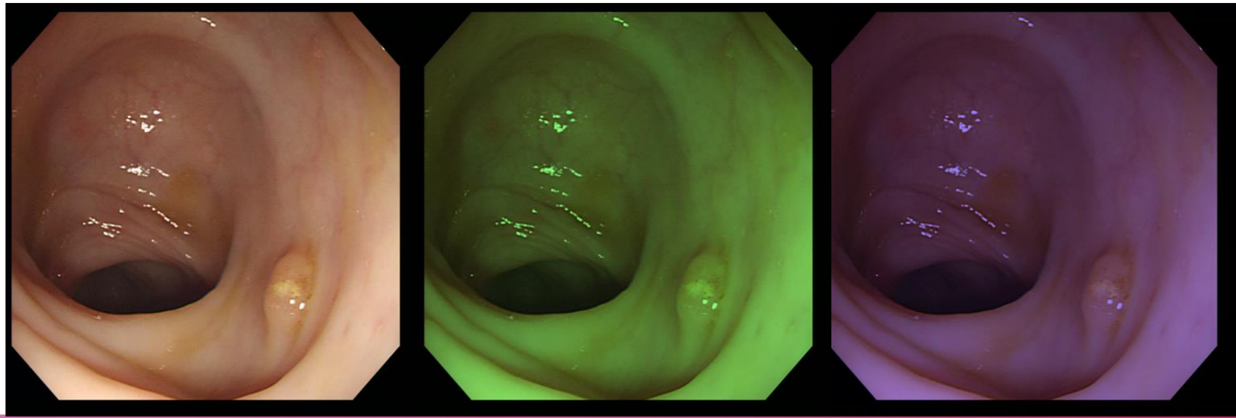
White Balance

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

Original Green Purple



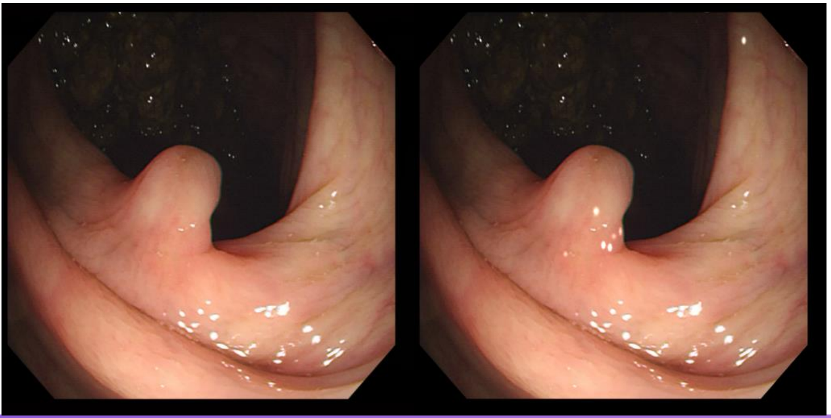
Original Green Purple



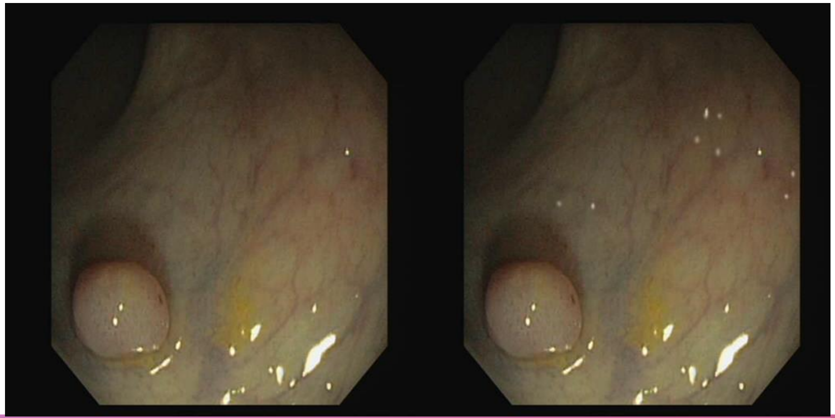
Specularity

- 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.

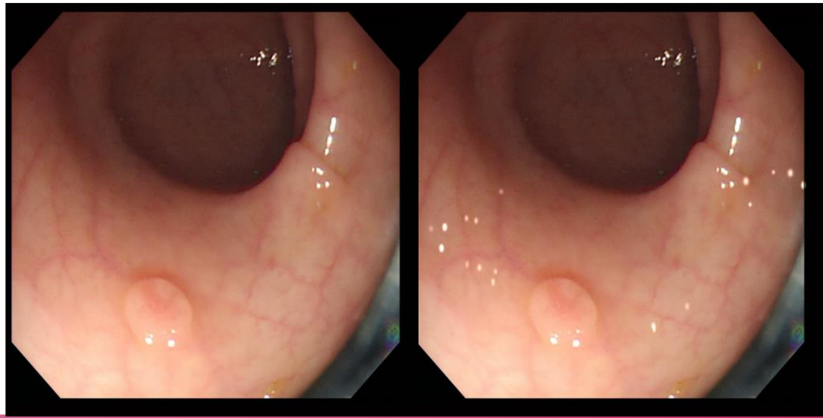
Original Specularity



Original Specularity



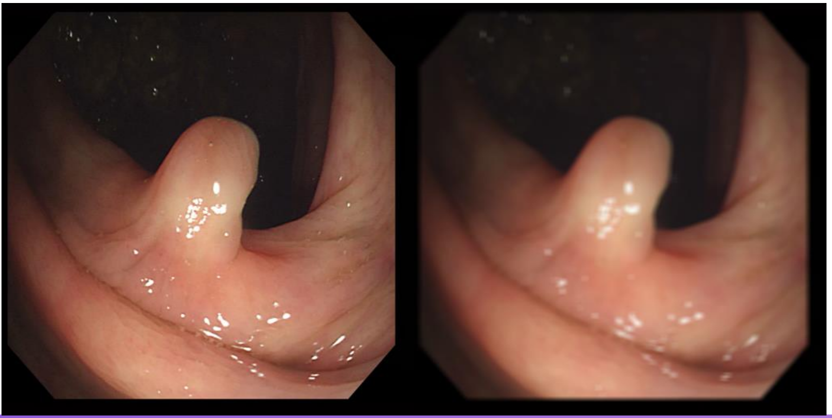
Original Specularity



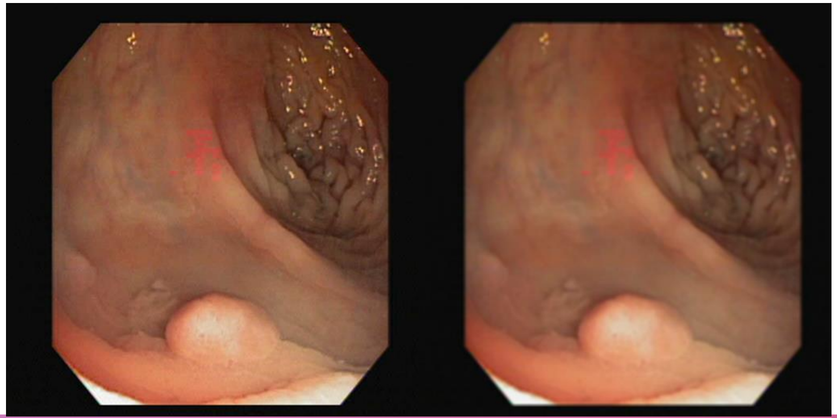
Motion Blur

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

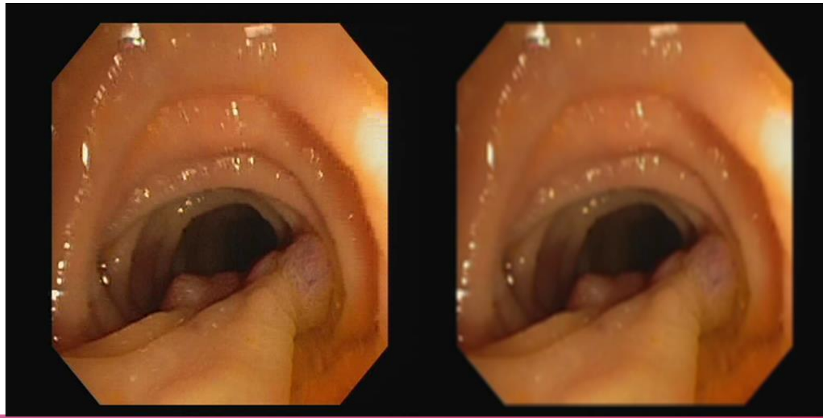
Original Blur



Original Blur



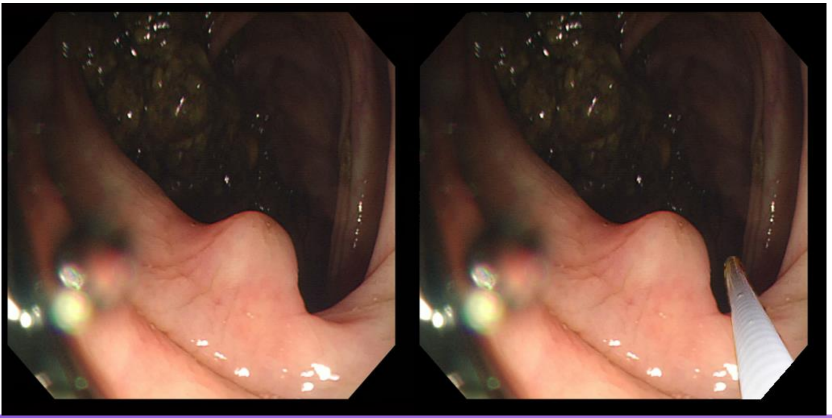
Original Blur



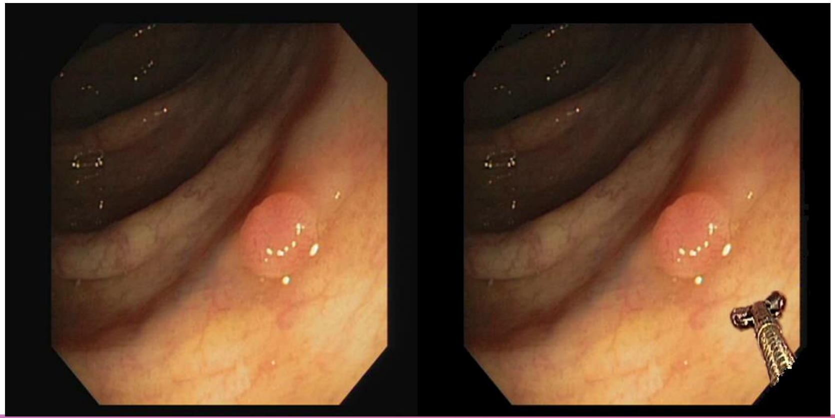
Instrument

- 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.

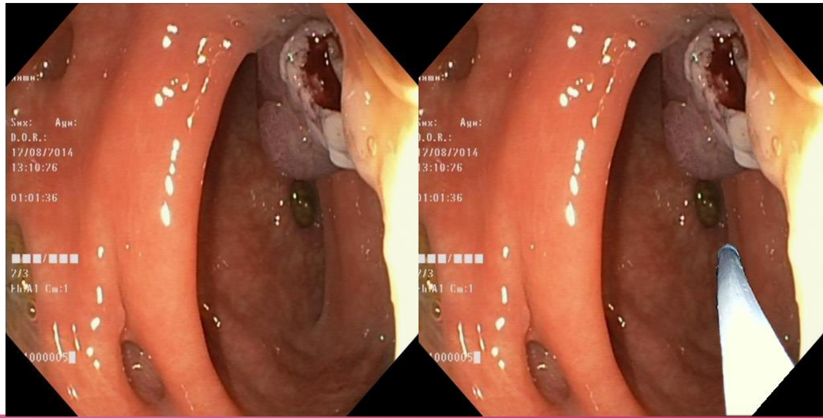
Original Instrument



Original Instrument



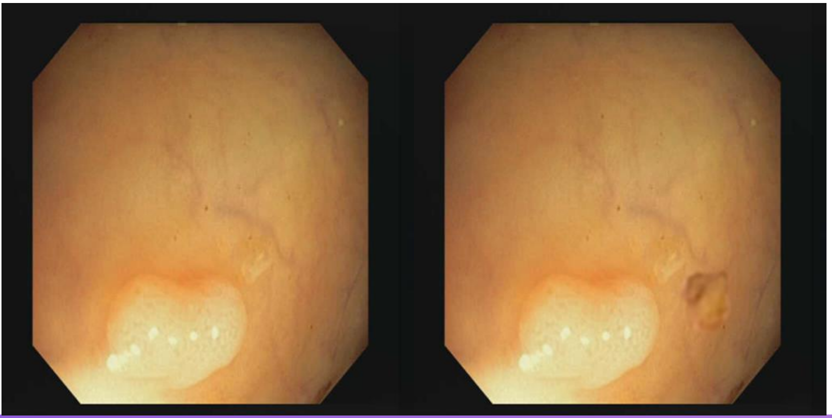
Original Instrument



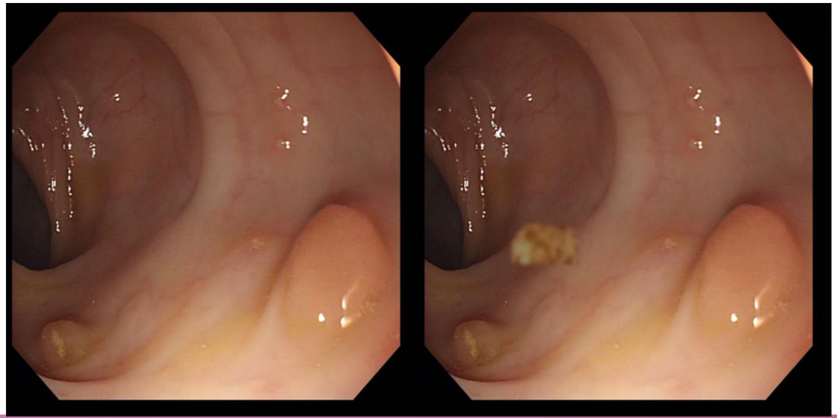
Feces

- 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.

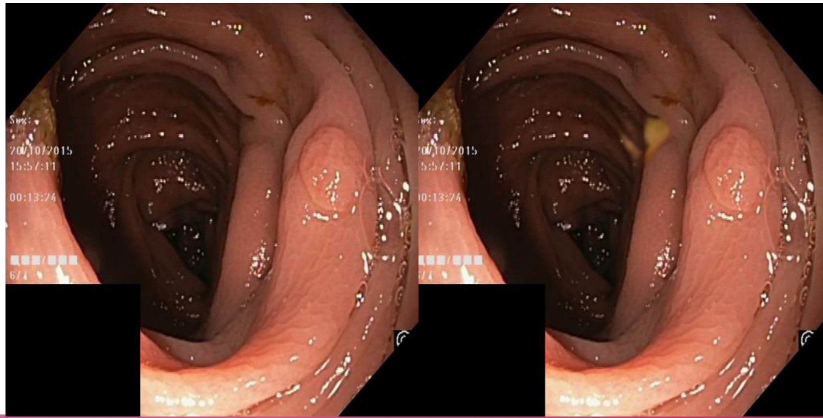
Original Feces



Original Feces



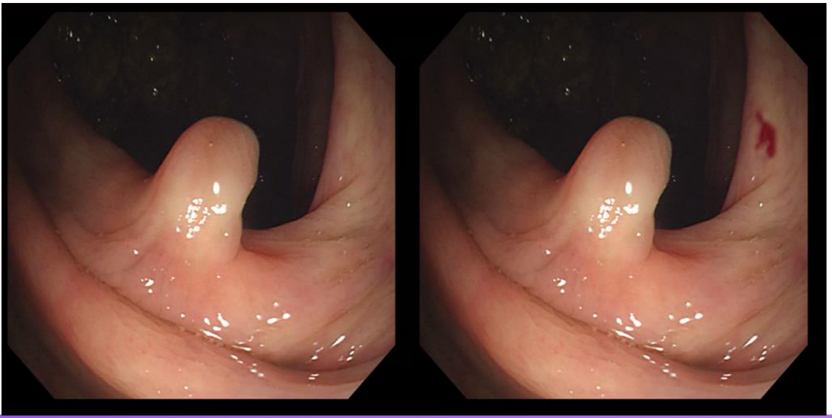
Original Feces



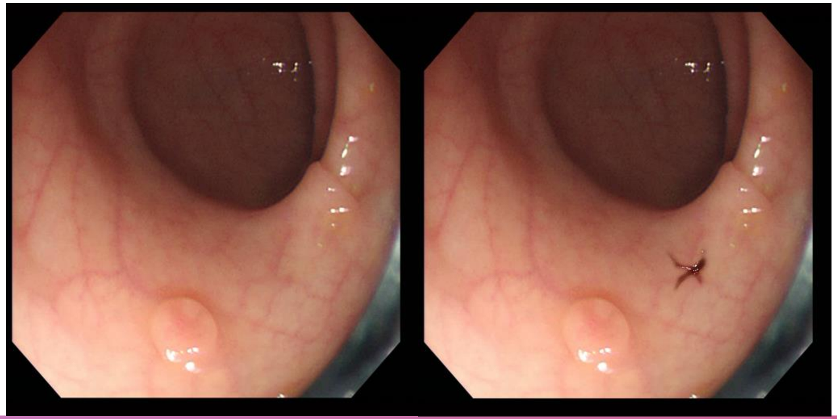
Blood

- 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.

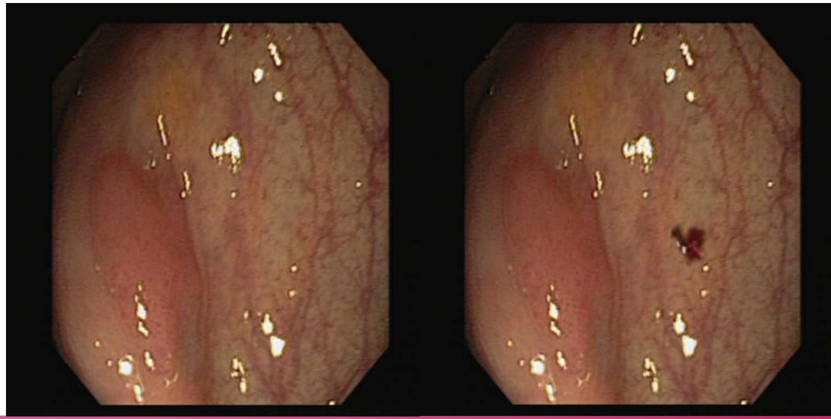
Original Blood



Original Blood



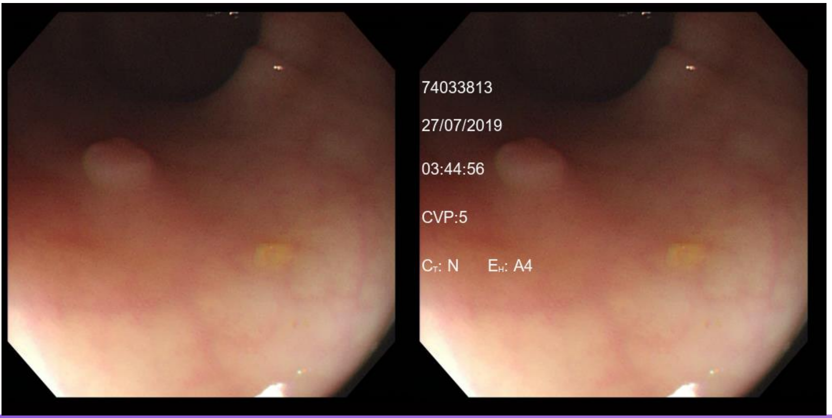
Original Blood



Text

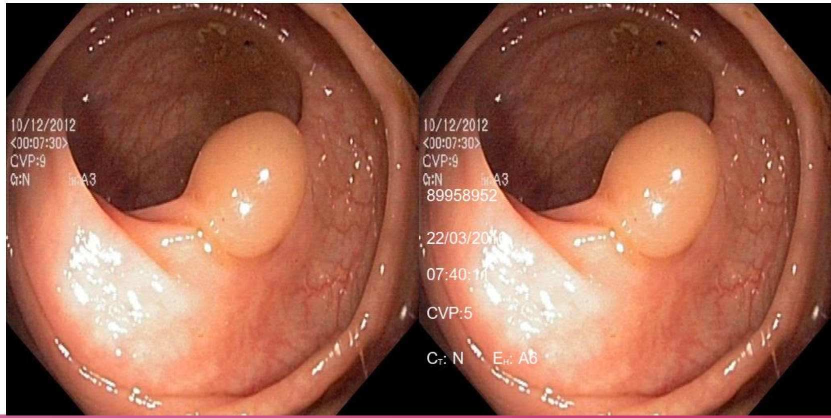
- Pattern in the text displayed on endoscopic images.
- Used ImageDraw method of PIL to generate text.

Original Text



Original Text

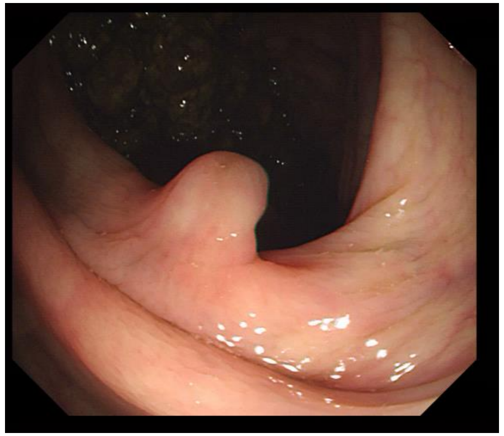
Original Text



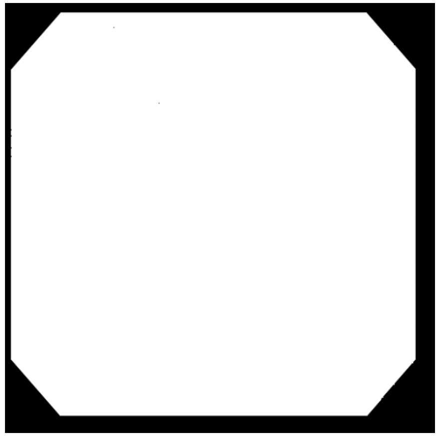
Data Pre-processing

- 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

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-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?

Experiment Settings

- Mainly utilize clinical **endoscopy images** to evaluate models.
- **Datasets**:
  - CVC-300 (60 images), CVC-ColonDB (380 images) mainly for the segmentation task. 440 seed images in total.
  - Additional ImageCLEF MEDVQA (182 images and 18 questions each) for VQA testing.
  - Wireless capsule endoscopy images from CAD-CAP and KID, with 600 seed images for testing in total.
- **Models under testing**:
  - Polyp **segmentation** models: PraNet, SANet, TGANet, SSFormer.
  - Multi-modal models for **Visual Question-Answering** (VQA): GPT-4V.
  - Gastrointestinal disease **classification**: AGDN, DSI-Net.

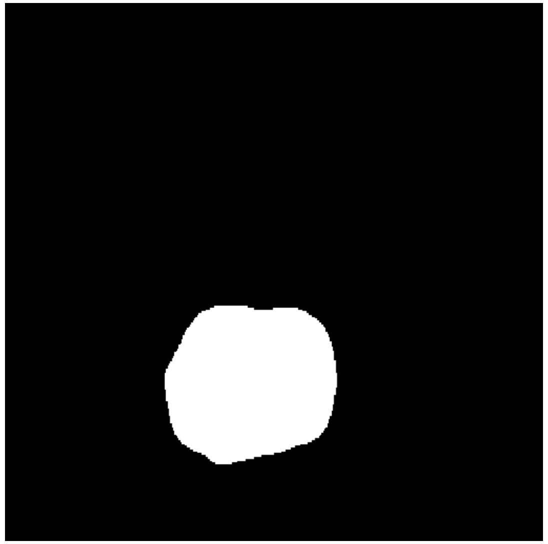
Evaluation-RQ1

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

- **Segmentation**
- Visual-Question Answering
- Classification

Segmentation

- Segmentation is the task to divide the image into different meaningful regions of interest.

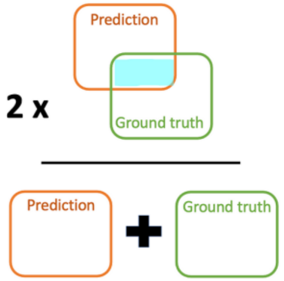


Evaluation Criteria

Measurement for segmentation task:

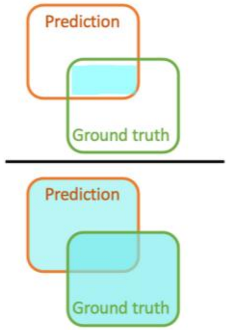
- Dice Score:

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



- 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

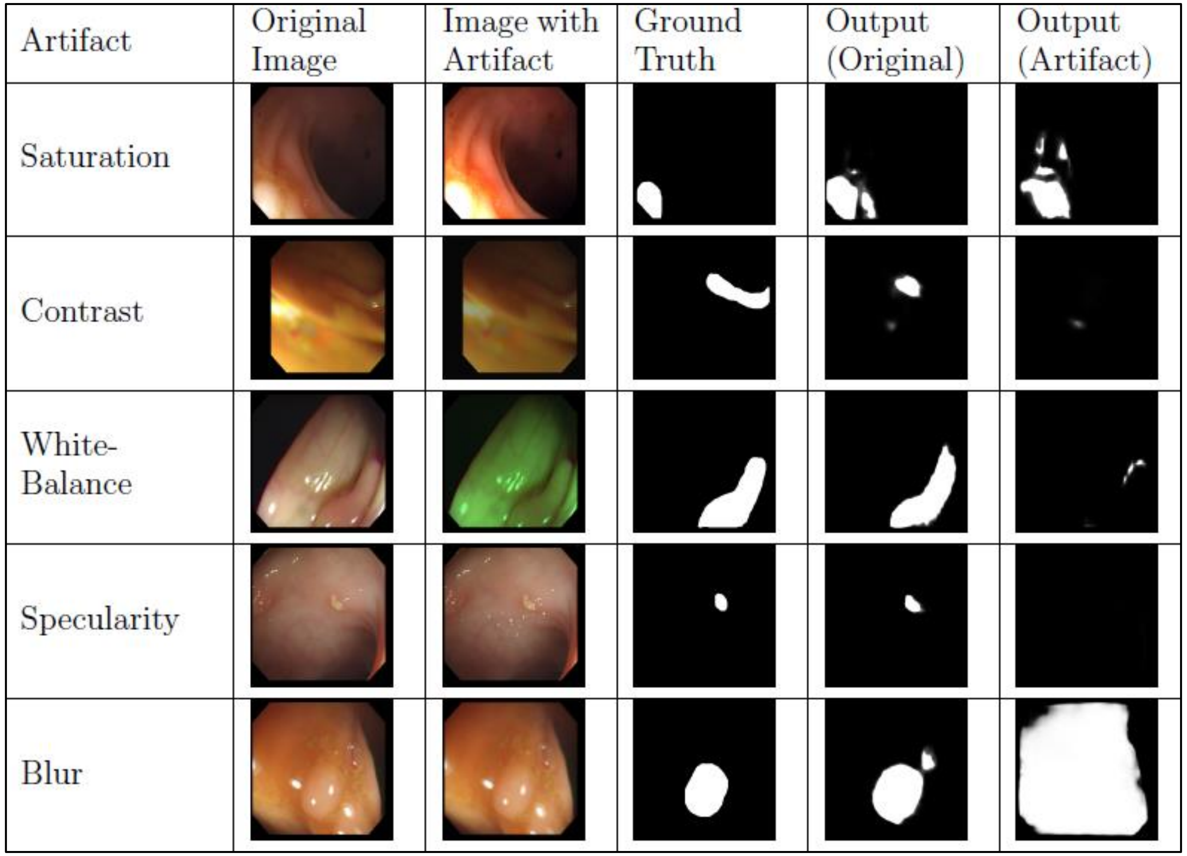
$$\frac{Original\ Score\ - Artifact\ Score}{Original\ Score} > t$$

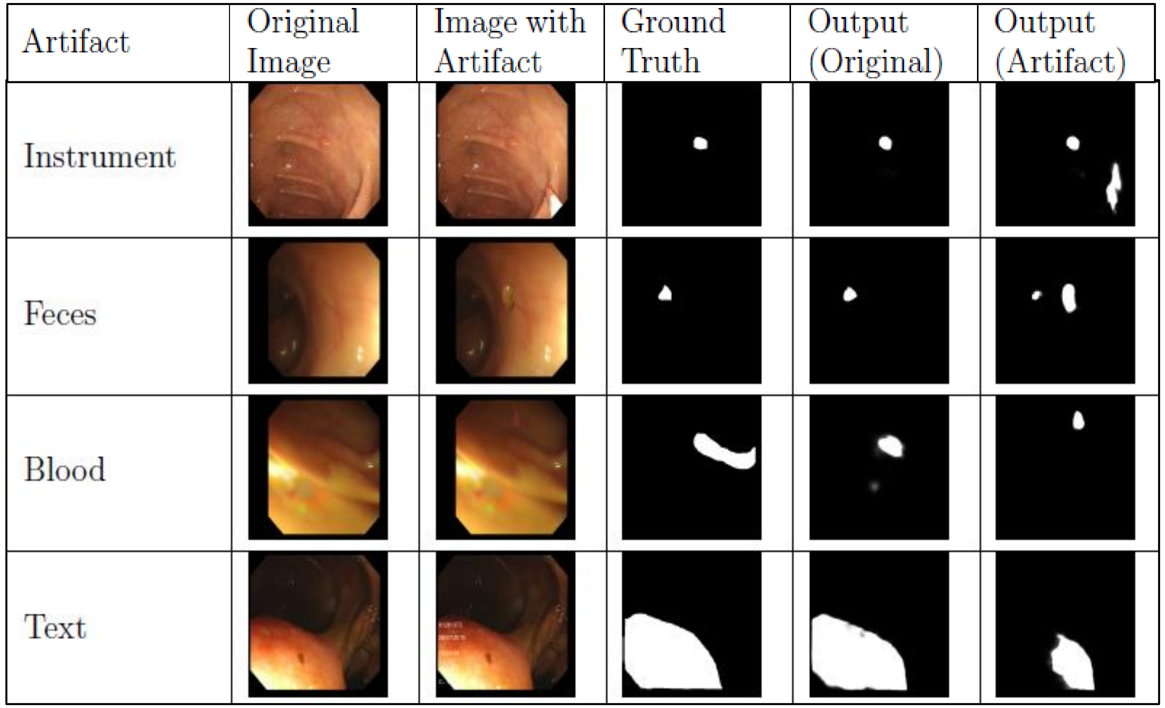
Error Finding Rate (EFR):

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

Illustration

- We evaluated four segmentation models (PraNet, SANet, TGANet, SSFormer) respectively.





Results

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

PraNet	CVC-300		CVC-ColonDB	
t=0.25	Dice	IoU	Dice	IoU
Blood	3.3	6.7	4.0	5.0
Feces	0.0	1.7	7.4	9.2
Instrument	6.7	11.7	12.1	14.0
Spot	1.7	1.7	3.2	4.2
Saturation	**8.3**	**13.3**	6.6	8.4
Contrast	1.7	5.0	4.7	6.1
White Balance	**8.3**	**13.3**	**19.8**	**22.7**
Blur	**8.3**	8.3	14.2	17.2
Text	0.0	0.0	5.0	5.8

PraNet: Overall EFR = 6.41%

SANet	CVC-300		CVC-ColonDB	
t=0.25	Dice	IoU	Dice	IoU
Blood	0.0	0.0	2.9	3.4
Feces	1.7	1.7	6.9	7.4
Instrument	1.7	1.7	5.5	5.8
Spot	0.0	0.0	4.2	4.5
Saturation	**5.0**	**6.7**	3.4	5.5
Contrast	0.0	0.0	3.4	4.0
White Balance	0.0	0.0	**10.8**	**14.0**
Blur	3.3	5.0	6.3	8.7
Text	0.0	0.0	5.5	5.8

SANet: Overall EFR = 3.37%

Results

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

TGANet	CVC-300		CVC-ColonDB	
t=0.25	Dice	IoU	Dice	IoU
Blood	16.7	20.0	23.9	29.2
Feces	13.3	25.0	13.9	18.2
Instrument	30.0	**46.7**	18.9	24.2
Spot	3.3	3.3	5.5	6.6
Saturation	16.7	18.3	21.8	24.7
Contrast	0.0	1.7	26.8	29.2
White Balance	**31.7**	38.3	**35.3**	**40.8**
Blur	28.3	31.7	9.7	11.8
Text	8.3	8.3	10.8	13.4

TGANet: Overall EFR = 17.49%

SSFormer	CVC-300		CVC-ColonDB	
t=0.25	Dice	IoU	Dice	IoU
Blood	3.3	3.3	5.0	5.3
Feces	0.0	0.0	7.6	8.2
Instrument	3.3	6.7	7.1	7.6
Spot	0.0	0.0	2.4	2.4
Saturation	**6.7**	**10.0**	2.6	4.5
Contrast	1.7	3.3	3.9	4.7
White Balance	3.3	5.0	**11.8**	**13.9**
Blur	0.0	1.7	3.4	3.4
Text	0.0	0.0	2.1	2.6

SSFormer: Overall EFR = 3.57%

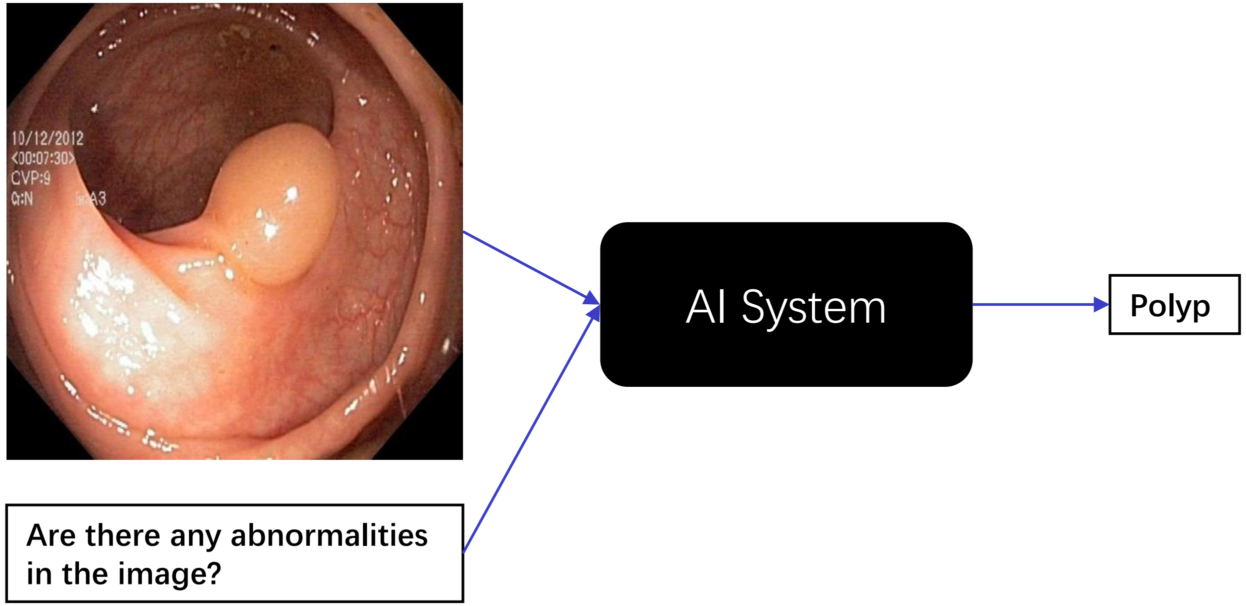
Evaluation-RQ1

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

- Segmentation
- **Visual-Question Answering**
- Classification

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

- Used the questions provided in the CLEF2023 MEDVQA Dataset.
- Includes "Yes/No" questions and simple questions regarding objects of interest within the medical images.
- Mainly conducted experiments on GPT-4V.

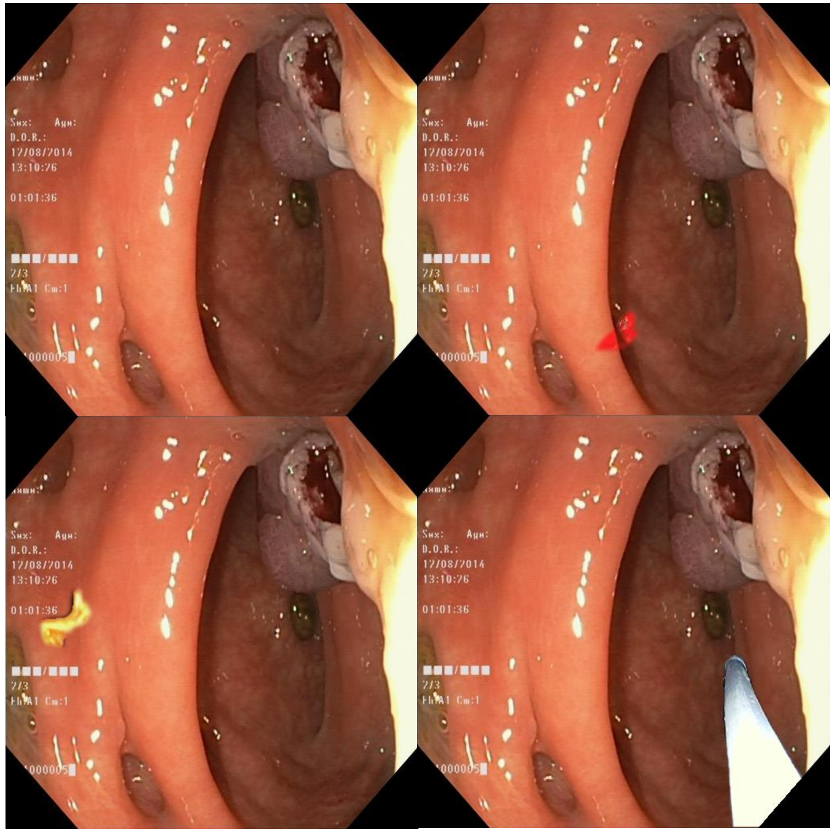
Question Number	Question
1	Are there any abnormalities in the image?
2	Are there any anatomical landmarks in the image?
3	Are there any instruments in the image?
4	Have all polyps been removed?
5	How many findings are present?
6	How many instruments are in the image?
7	How many polyps are in the image?
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?

Illustration

- Illustration on VQA testing case

Original

Blood



Feces

Instrument

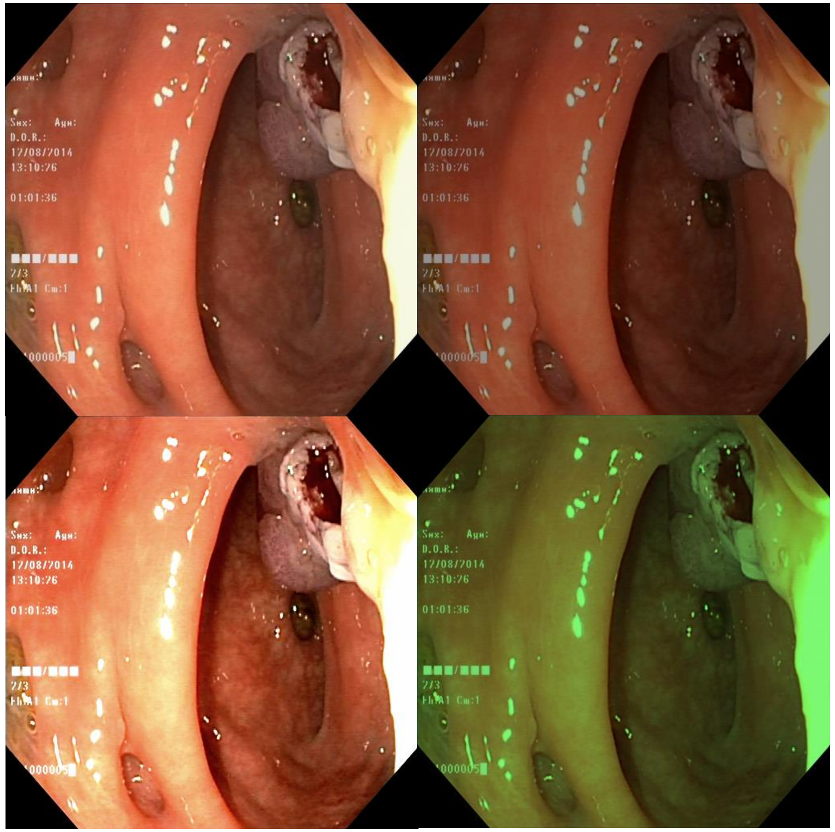
Question	Ground Truth	Original	Blood	Feces	Instrument
Are there any abnormalities in the image?	Polyp	Polyp	Bleeding	Feces	Polyp
Are there any anatomical landmarks in the image?	No	No	No	Yes	Yes
Are there any instruments in the image?	No	No	No	No	Yes
Have all polyps been removed?	No	No	Not relevant	Not 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

Illustration

- Illustration on VQA testing case

Original

Contrast



Saturation

White Balance

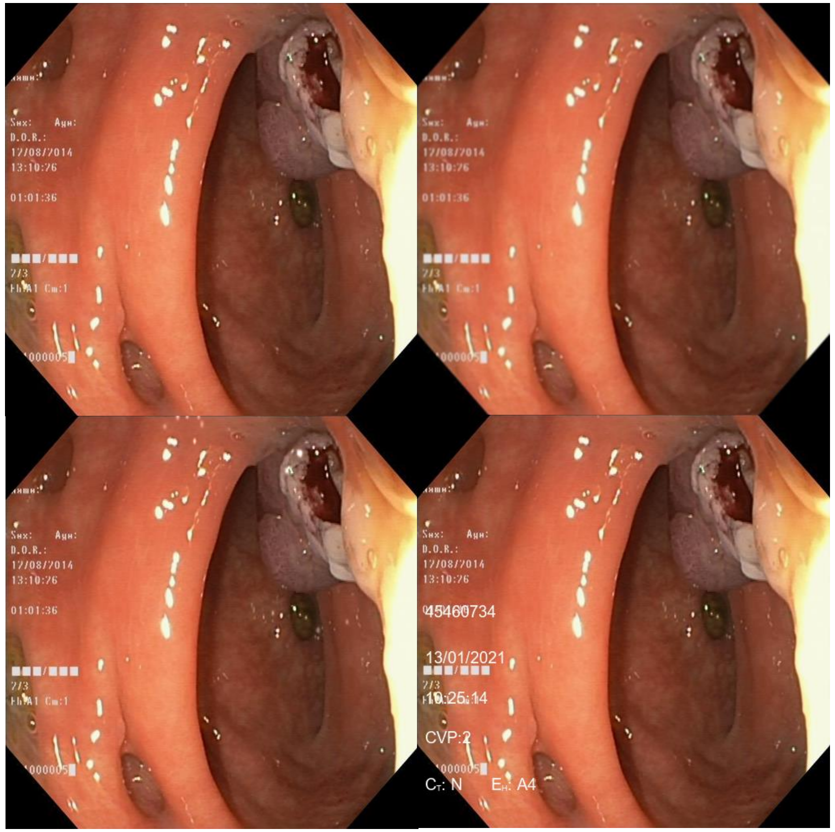
Question	Ground Truth	Original	Contrast	Saturation	White Balance
Are there any abnormalities in the image?	Polyp	Polyp	Polyp	Polyp	Polyp
Are there any anatomical landmarks in the image?	No	No	Yes	Yes	Yes
Are there any instruments in the image?	No	No	No	No	No
Have all polyps been removed?	No	No	No	No	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

Illustration

- 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

Results

- The VQA testing result on GPT-4V.
- The model is quite robust.

GPT-4V	Original	Average Perturbation	Difference (Original - Average)
Are there any abnormalities in the image?	0.888	0.861	0.027
Are there any anatomical landmarks in the image?	0.341	0.347	-0.006
Are there any instruments in the image?	0.947	0.873	0.074
Have all polyps been removed?	0.284	0.282	0.002
How many findings are present?	0.835	0.816	0.019
How many instrumnets are in the image?	0.963	0.93	0.033
How many polyps are in the image?	0.79	0.763	0.027
Is there a green/black box artefact?	0.624	0.636	-0.012
Is there text?	0.98	0.963	0.017
Is this finding easy to detect?	0.594	0.589	0.005
What color is the abnormality?	0.465	0.406	0.059
What color is the anatomical landmark?	0.369	0.388	-0.019
What is the size of the polyp?	0.21	0.206	0.004
What type of polyp is present?	0.176	0.163	0.013
What type of procedure is the image taken from?	0.976	0.964	0.012
Where in the image is the abnormality?	0.653	0.644	0.009
Where in the image is the anatomical landmark?	0.365	0.386	-0.021
Where in the image is the instrument?	0.924	0.896	0.028
Average	0.632	0.617	0.015

Overall score before deletion

Results

- The VQA testing result on GPT-4V.
- We found that some questions are ambiguous and not related to our task.

GPT-4V	Original	Average Perturbation	Difference (Original - Average)
Are there any abnormalities in the image?	0.888	0.861	0.027
Are there any instruments in the image?	0.947	0.873	0.074
Have all polyps been removed?	0.284	0.282	0.002
How many findings are present?	0.835	0.816	0.019
How many instrumnets are in the image?	0.963	0.93	0.033
How many polyps are in the image?	0.79	0.763	0.027
Is there a green/black box artefact?	0.624	0.636	-0.012
Is there text?	0.98	0.963	0.017
What color is the abnormality?	0.465	0.406	0.059
What is the size of the polyp?	0.21	0.206	0.004
What type of polyp is present?	0.176	0.163	0.013
What type of procedure is the image taken from?	0.976	0.964	0.012
Where in the image is the abnormality?	0.653	0.644	0.009
Where in the image is the instrument?	0.924	0.896	0.028
Average	0.694	0.672	0.022

Overall score after deletion

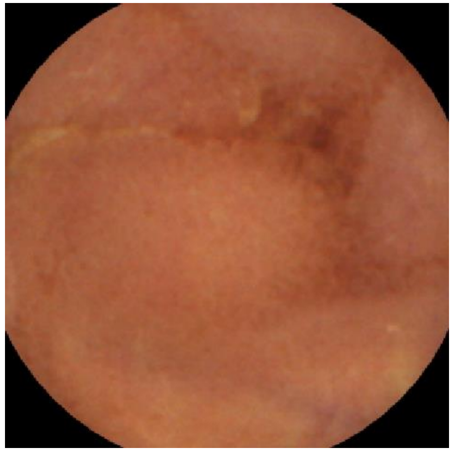
Evaluation-RQ1

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

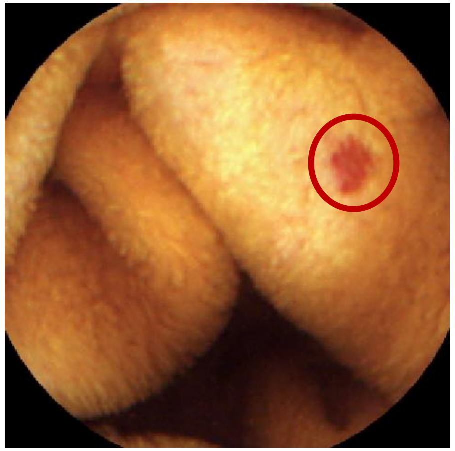
- Segmentation
- Visual-Question Answering
- **Classification**

Classification

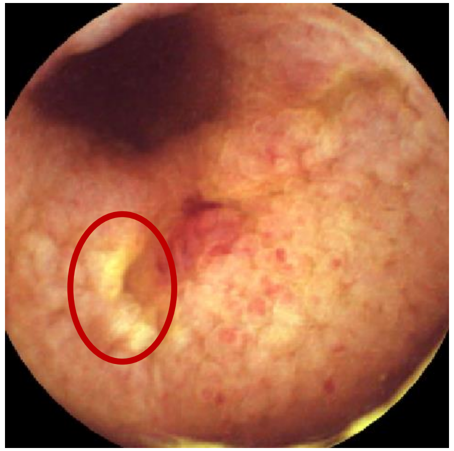
- A medical image analysis technique that involves classifying medical images into different categories based on the type of image or the presence of specific structures or diseases.
- Our Task: Classify wireless capsule endoscopy (WCE) images into three types – Normal, Vascular Lesion, Inflammatory



Normal



Vascular Lesion



Inflammatory

Evaluation Criteria

Measurement for classification task:

- **Accuracy**: the proportion of accurately classified samples to the total number of test cases
- **F1 Score**: a weighted harmonic mean of Precision and Recall normalized between 0 and 1

$$F1\ Score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

- **Cohen's Kappa Score**: measures the proximity of the predicted classes to the actual classes when compared to a random classification

Results

- The performance of AGDN and DSI-Net on different perturbations, respectively.
- We can find a maximum decrease in the accuracy of **0.229** and **0.144** in the two models respectively.

Artifact	Accuracy	Cohen's Kappa	F1 Score
Original	0.893	0.836	0.893
Blur	0.702	0.517	0.660
Contrast	0.747	0.602	0.735
Feces	0.797	0.682	0.790
Instrument	0.852	0.773	0.852
Saturation	0.685	0.516	0.682
Spot	0.817	0.712	0.811
Text	0.828	0.733	0.825
White Balance	0.532	0.226	0.463
Average	0.745	0.595	0.727

AGDN

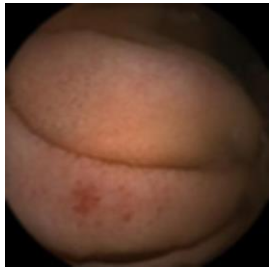
Artifact	Accuracy	Cohen's Kappa	F1 Score
Original	0.940	0.908	0.940
Blur	0.897	0.841	0.896
Contrast	0.883	0.823	0.884
Feces	0.907	0.858	0.907
Instrument	0.897	0.843	0.897
Saturation	0.755	0.635	0.757
Spot	0.932	0.895	0.931
Text	0.908	0.859	0.908
White Balance	0.728	0.574	0.711
Average	0.863	0.791	0.861

DSI-Net

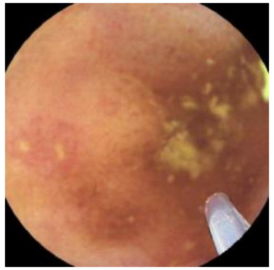
Perturbation Analysis

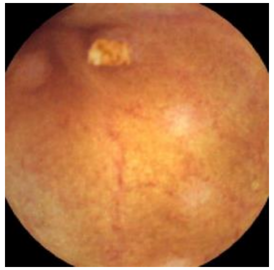
- Lighting conditions (e.g. white balance, saturation) can trigger most errors.
  - Possible explanation: 1. Edges of critical regions become vague 2. Sensitive to color

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



- Object-related perturbations resulted in some misleading cases.
  - Possible explanation: 1. Unseen elements (e.g., Instruments) in the training data 2. Feces resemble some objects of interest





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 obtained up to 17.49% EFR when testing on segmentation models, qualitatively affected VQA models' performances, and reduced the accuracy of classification models for up to 16.6% in average, indicating 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?

Evaluation-RQ2

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

- **Segmentation**
- Classification

Further Training Settings

- We have further trained the models with perturbed images synthesized by MedTest to improve the performances of those models.

- Datasets Construction:
    - CVC-ClinicDB (612 images) and Kvasir (1000 images)
    - Besides the original training set, we also randomly selected an equal number of perturbed images generated from the original training set.

- We used the same training settings as stated in their original papers.

Results

- We used Dice and IoU scores directly for comparison.
- “Blur” has an over 10% increase across both datasets and metrics.

PraNet	CVC-300				CVC-ColonDB			
	Dice Score		IoU Score		Dice Score		IoU Score	
	Before	After	Before	After	Before	After	Before	After
Original	0.86	**0.889**	0.777	**0.817**	0.696	**0.701**	0.619	**0.629**
Saturation	0.818	**0.878**	0.735	**0.806**	0.677	**0.711**	0.598	**0.636**
White Balance	0.815	**0.879**	0.74	**0.805**	0.624	**0.709**	0.551	**0.634**
Contrast	0.861	**0.887**	0.777	**0.815**	0.692	**0.707**	0.618	**0.633**
Spot	0.849	**0.885**	0.764	**0.812**	0.694	**0.702**	0.615	**0.628**
Blur	0.706	**0.872**	0.619	**0.794**	0.582	**0.695**	0.491	**0.618**
Text	0.744	**0.819**	0.659	**0.746**	0.629	**0.663**	0.554	**0.593**
Instrument	0.812	**0.879**	0.717	**0.805**	0.653	**0.699**	0.571	**0.626**
Blood	0.843	**0.891**	0.751	**0.821**	0.678	**0.696**	0.599	**0.624**
Feces	0.838	**0.878**	0.75	**0.804**	0.674	**0.691**	0.592	**0.619**
Average	0.815	**0.876**	0.729	**0.803**	0.66	**0.697**	0.581	**0.624**

PraNet: average improvement 4.9%

Results

- We used Dice and IoU score directly for comparison.
- “Text” achieving up to a 6.6% increase in the IoU score on CVC-300.

SANet	CVC-300				CVC-ColonDB			
	Dice Score		IoU Score		Dice Score		IoU Score	
	Before	After	Before	After	Before	After	Before	After
Original	0.898	**0.904**	0.828	**0.836**	0.757	**0.763**	0.677	**0.69**
Saturation	0.879	**0.882**	0.807	**0.816**	**0.766**	0.757	0.683	0.683
White Balance	0.877	**0.889**	0.805	**0.822**	0.738	**0.76**	0.655	**0.687**
Contrast	**0.898**	0.897	0.827	**0.829**	0.754	**0.765**	0.673	**0.691**
Spot	0.899	**0.904**	0.828	**0.837**	0.754	**0.76**	0.675	**0.686**
Blur	0.851	**0.899**	0.773	**0.831**	0.735	**0.773**	0.646	**0.697**
Text	0.803	**0.864**	0.727	**0.793**	0.694	**0.749**	0.621	**0.674**
Instrument	0.898	**0.903**	0.829	**0.836**	0.747	**0.749**	0.668	**0.679**
Blood	0.899	**0.903**	0.829	**0.835**	0.753	**0.762**	0.675	**0.689**
Feces	0.901	**0.904**	0.831	**0.837**	0.743	**0.759**	0.665	**0.687**
Average	0.88	**0.895**	0.808	**0.827**	0.744	**0.76**	0.664	**0.686**

SANet: average improvement 1.6%

Results

- We used Dice and IoU score directly for comparison.
- “White Balance” achieving up to a 3.7% increase in the IoU score on CVC-ColonDB

SSFormer	CVC-300				CVC-ColonDB			
	Dice Score		IoU Score		Dice Score		IoU Score	
	Before	After	Before	After	Before	After	Before	After
Original	0.891	0.891	0.825	**0.827**	0.774	0.774	0.698	**0.700**
Saturation	0.841	**0.876**	0.779	**0.806**	0.778	0.778	0.699	**0.702**
White Balance	**0.880**	0.874	**0.813**	0.811	0.731	**0.764**	0.656	**0.693**
Contrast	**0.883**	0.882	0.817	**0.817**	0.765	**0.774**	0.689	**0.700**
Spot	0.892	0.892	0.826	**0.828**	0.770	**0.775**	0.695	**0.701**
Blur	0.883	**0.893**	0.813	**0.825**	0.766	0.766	0.690	**0.693**
Text	0.892	**0.893**	0.825	**0.830**	0.768	**0.769**	0.691	**0.696**
Instrument	0.872	**0.886**	0.800	**0.820**	0.747	**0.769**	0.672	**0.695**
Blood	0.877	**0.891**	0.809	**0.827**	0.760	**0.770**	0.684	**0.697**
Feces	0.898	**0.899**	0.831	**0.835**	0.753	**0.759**	0.679	**0.687**
Average	0.881	**0.888**	0.814	**0.823**	0.761	**0.770**	0.684	**0.696**

SSFormer: average improvement 0.8%

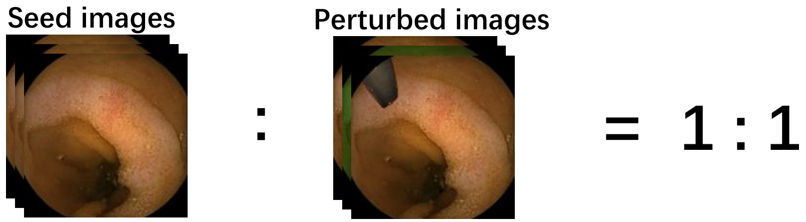
Evaluation-RQ2

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

- Segmentation
- **Classification**

Further Training Settings

- **Dataset construction**:
  - Excluded blood perturbation in the dataset construction
  - 2422 images in the training set + randomly select $\frac{1}{8}$ of the total images in each perturbation generated from the training set



- In general, we follow the original model setting in the implementation
- Minor adjustments on parameters such as the learning rate and batch size
- Performed train-validation split (0.85 : 0.15) on our customized data and used the validation performance to select checkpoints

Results

- Classification Performance on AGDN model:

Artifact	Initial Accuracy	Enhanced Accuracy	Difference
Original	0.893	0.885	-0.008
Blur	0.702	0.808	**+0.106**
Contrast	0.747	0.812	**+0.065**
Feces	0.797	0.860	**+0.063**
Instrument	0.852	0.852	0.000
Saturation	0.685	0.770	**+0.085**
Spot	0.817	0.873	**+0.056**
Text	0.828	0.858	**+0.030**
White Balance	0.532	0.673	**+0.141**
Average	0.761	0.821	**+0.060**

AGDN: The average improvement in accuracy score is 6%

Results

- Classification Performance on DSI-Net model:

Artifact	**Initial Accuracy**	**Enhanced Accuracy**	**Difference**
Original	0.940	0.947	**+0.007**
Blur	0.897	0.918	**+0.021**
Contrast	0.883	0.917	**+0.034**
Feces	0.907	0.937	**+0.030**
Instrument	0.897	0.928	**+0.031**
Saturation	0.755	0.835	**+0.080**
Spot	0.932	0.942	**+0.010**
Text	0.908	0.908	0.000
White Balance	0.728	0.848	**+0.120**
Average	0.872	0.909	**+0.037**

DSI-Net: The average improvement in accuracy score is 3.7%

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: Test cases generated by MedTest can be leveraged to construct our customized training dataset and effectively improve the robustness of academic medical image diagnosis models through further training on both segmentation and classification tasks.

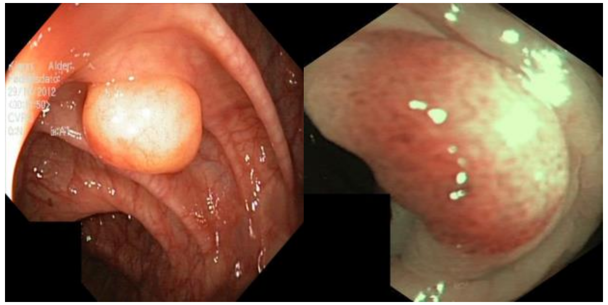
Evaluation-RQ3

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

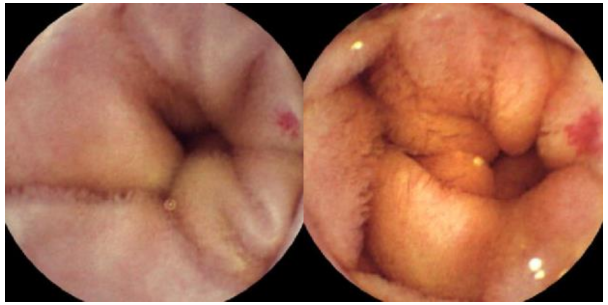
- RQ1: Is our method effective in identifying incorrect outputs produced by medical image diagnosis software and algorithms?
- RQ2: Can the test cases generated by our method be utilized to enhance the performance of medical image diagnosis software?
- **RQ3: What are the various factors that influence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

External Factors and Influences

- Divergence in Image Structure and Overlay
  - Hard to decide on suitable places for object-related perturbations
- Medical Landmarks Characteristics
  - Need to avoid affecting the original medical characteristics
- Lighting Conditions
  - Require targeting methods to keep the lighting condition consistent and realistic



Samples of various polyp shapes



Samples of vascular lesions (bleeding)

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: The performance of MedTest can potentially be affected by the above - proposed factors, including image structure, medical landmark characteristics, and ambient lighting conditions. We have considered these factors in the design of MedTest and tried to mitigate the negative effect to the greatest extent in our implementation.

Future Work

- Expand our testing objectives to other large language models and multi-modal models specially on medical diagnosis, e.g., Gemini and Med-PaLM.

Conclusion

- Targeting the important inter-discipline of AI and medicine, we designed a comprehensive metamorphic testing paradigm, MedTest, to comprehensively evaluate models and software on medical imaging tasks.

- With our clinical-equivalent perturbations, our method was proved to effectively identify potential model errors and showed potential in assisting to improve model performance.

- Future work focuses on expanding the testing objectives of MedTest, especially in medical multi-modal 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).

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

[10] Acosta, Julián N., et al. "Multimodal biomedical AI." *Nature Medicine* 28.9 (2022): 1773-1784.

[11] Xing, Xiaohan, Yixuan Yuan, and Max Q-H. Meng. "Zoom in lesions for better diagnosis: Attention guided deformation network for wce image classification." *IEEE Transactions on Medical Imaging* 39.12 (2020): 4047-4059.

[12] Zhu, Meilu, Zhen Chen, and Yixuan Yuan. "DSI-Net: Deep synergistic interaction network for joint classification and segmentation with endoscope images." *IEEE Transactions on Medical Imaging* 40.12 (2021): 3315-3325.

[13] Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

[14] Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." *arXiv preprint arXiv:2312.11805* (2023).