Using Synthetic and Pseudosynthetic Data to Enhance Polyp Detection in Future AI-assisted Endoscopy Frameworks

Andrei-Constantin IOANOVICI[1], Marius-Ștefan MĂRUȘTERI[1]*, Andrei Marian FEIER[2], Vasile Florin POPESCU[3], Irina IOANOVICI[4], Daniela-Ecaterina DOBRU[5]

[1] Department M2 - Complementary Functional Sciences, Medical Informatics and Biostatistics, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, 1 Gheorghe Marinescu St., Targu Mures, 540142, Romania
andrei.ioanovici@umfst.ro, marius.marusteri@umfst.ro (**Corresponding author*)

[2] Department M4 - Clinical Sciences, Orthopedics and Traumatology I, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, 1 Gheorghe Marinescu St., Targu Mures, 540139, Romania
andrei.feier@umfst.ro

[3] "Carol I" National Defense University, Sos. Panduri no. 68-72, Bucharest, 050662, Romania
popescuveve@gmail.com

[4] Emergency Clinical County Hospital of Targu-Mures, Allergology and Immunology Unit, 50 Gheorghe Marinescu St., Targu Mures, 540136, Romania
irina.naumof@gmail.com

[5] Department M4 Clinical Sciences, Gastroenterology Medical VII, George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures, 1 Gheorghe Marinescu St., Targu Mures, 540103, Romania
danidobru@gmail.com

Abstract: Colorectal cancer (CRC) incidence can be reduced through the early detection and removal of precancerous polyps. Artificial intelligence (AI), especially deep learning, enhances polyp detection during colonoscopy but it often faces limitations from small medical imaging datasets. This study investigates whether synthetic and pseudosynthetic data-augmented images derived from original datasets can improve AI accuracy in polyp detection. Pseudosynthetic data, uniquely derived through augmentation techniques such as flipping, rotation, and contrast adjustment, simulates multiple endoscopic examinations of the same patients without subjecting them to repeated invasive procedures, while enabling the traceability of the original clinical data. A modified U-Net was trained on various combinations of real, synthetic (CycleGAN and diffusion-based), and pseudosynthetic datasets across ten experimental setups and externally validated on the CVC-Colon-DB dataset (including 612 images). The combination of real and pseudosynthetic data provided the highest model performance (a Dice coefficient of 0.7638, a precision of 0.8979, a recall of 0.7535, and a F1 score of 0. 0.7797). To that, when the proposed model employed diffusion-based synthetic data it performed better than when using CycleGAN-generated data, which demonstrated its superior generalization capability in the former case (with a precision of 0.7488, a recall of 0.6695 and a F1 score of 0.8987). The obtained results show that pseudosynthetic data alone can significantly improve the generalization capability of the employed model in comparison with simply real data. These findings confirm that the augmented and synthetic datasets are valuable tools for enhancing a model's performance and addressing ethical concerns in AI-assisted diagnostics.

Keywords: Colon polyps, Synthetic data, Polyp detection, Polyp segmentation, Colorectal cancer.

1. Introduction

Colorectal cancer (CRC) is a prevalent malignancy globally (Morgan et al., 2023), however, its incidence can be mitigated through risk factor modification and the removal of precancerous lesions (Morrow & Greenwald, 2022; Wilhelmi et al., 2021; Sullivan et al., 2022). Incorporating Artificial Intelligence (AI), especially deep learning, into the digestive endoscopy significantly advances early CRC diagnosis and treatment, particularly with regard to polyp detection. AI systems enhance diagnostic capabilities by analyzing large datasets of annotated images using deep neural networks (Ahmad et al., 2019).

High adenoma miss rates during endoscopy remain critical, as various studies indicate that many polyps are overlooked even by skilled endoscopists. While additional colonoscopies might reduce miss rates, conducting repeated procedures is impractical due to patient risks, discomfort, and the strain on healthcare resources (Jiang et al., 2023; Herszényi, 2019) AI-assisted colonoscopy has shown its potential to enhance polyp detection rates by compensating for human error and variability (Barua et al., 2021; Shao et al., 2022). However, the development of these AI tools involves challenges such as ethical considerations regarding patient privacy, consent, and adherence to data protection laws when using clinical data (Williamson & Prybutok, 2024).

In training recent AI image recognition algorithms (Ronneberger, Fischer & Brox, 2015) real clinical

https://doi.org/10.24846/v34i2y202507

datasets as well as synthetic datasets generated through advanced techniques were utilized, including GANs (Zhu et al., 2017; Isola et al., 2017) and diffusion models (Dorjsembe, Pao, & Xiao, 2024). Furthermore, an innovative concept termed pseudosynthetic data was introduced. Although traditional data augmentation techniques such as rotations, flips, and contrast adjustments are commonly used to enhance dataset variability, pseudosynthetic data extends beyond these methods by introducing clinically meaningful simulations. Specifically, it simulates realistic clinical scenarios by presenting the same polyp from multiple perspectives under varying conditions such as lighting, angle, and focus, mimicking repeated endoscopic examinations. Pseudosynthetic data maintains traceability to its original clinical source, allowing direct linkage between the augmented images and real patient data. This explicit connection supports validation, transparency, and reproducibility, distinguishing pseudosynthetic data-based methods clearly from conventional augmentation methods, where augmented images often lack clinical relevance or identifiable connections to the original data.

By augmenting the existing datasets in this clinically relevant manner, pseudosynthetic data simulates repeated endoscopic procedures without subjecting patients to multiple invasive colonoscopies within a short timeframe - a practice generally not recommended (van Liere et al., 2023; Chen et al., 2022). This approach addresses ethical and practical concerns associated with repetitive clinical procedures, enhances data complexity and diversity, and significantly improves model generalization. Additionally, pseudosynthetic data can be extended to other endoscopic procedures and various medical data types, including numerical datasets.

While deep learning in medical imaging is well-studied, the specific role of augmented and synthetic data in enhancing diagnostic accuracy for colon polyps remains underexplored. This study evaluates the impact of synthetic data and pseudosynthetic data – a term for data that simulates image variability as if obtained from multiple endoscopies – on enhancing the diagnostic accuracy of deep learning models for colon polyp detection. Specifically, this paper investigates whether pseudosynthetic and synthetic data effectively address challenges related to data scarcity, lack of diversity in real-world datasets, and ethical concerns regarding patient privacy in AI-assisted diagnostics. Furthermore, it is assessed whether synthetic data generated via diffusion algorithms demonstrates a superior performance in comparison with GAN-generated data.

The remainder of the paper is as follows. Section 2 describes the utilised datasets, it outlines the training of a U-Net, and presents the experiments which were carried out using different mixes of real, pseudosynthetic and synthetic data. Section 3 presents the quantitative outcomes of seven distinct training experiments. Further on, Section 4 addresses the challenge posed by limited, ethically obtainable colonoscopy images and evaluates how data augmentation and synthesis mitigate that constraint. Finally, Section 5 synthesizes the insights of this paper, namely that enlarged, diversity-rich training datasets resolve data-scarcity and privacy constraints, the models trained with pseudosynthetic images achieve the highest improvements and diffusion-based synthesis surpasses GAN-derived datasets.

2. Materials and Methods

2.1 Data Sources, Preprocessing and Augmentation

Experiments were conducted using real and synthetic polyp datasets, along with pseudosynthetic datasets as detailed below.

The Kvasir-SEG dataset contains 1,000 polyp images with the corresponding ground truth masks, with resolutions ranging from 332×487 to 1920×1072 pixels (Jha et al., 2019) The PolypGen dataset includes colonoscopy images from six centers, involving over 300 patients, totaling 3,762 annotated polyp labels verified by six senior gastroenterologists, and features both single-frame and sequential data (Ali et al., 2023).

Synth-Colon was employed, too, a synthetic dataset comprising 20,017 realistic images generated using CycleGAN in conjunction with the Kvasir dataset (Zhu et al., 2017; Isola et al., 2017) Additionally, synthetic datasets of 20,000 polyp images were generated using a diffusion-based semantic polyp synthesis method - Denoising Diffusion Probabilistic Models (DDPM) guided by 5,000 masks. These synthetic images augment both the volume and diversity of the training data, aiding in the development of robust and generalizable models (Dorjsembe,

Pao & Xiao, 2024; Ho, Jain & Abbeel, 2020) A flowchart representing the synthetic image generation is displayed in Figure 1.

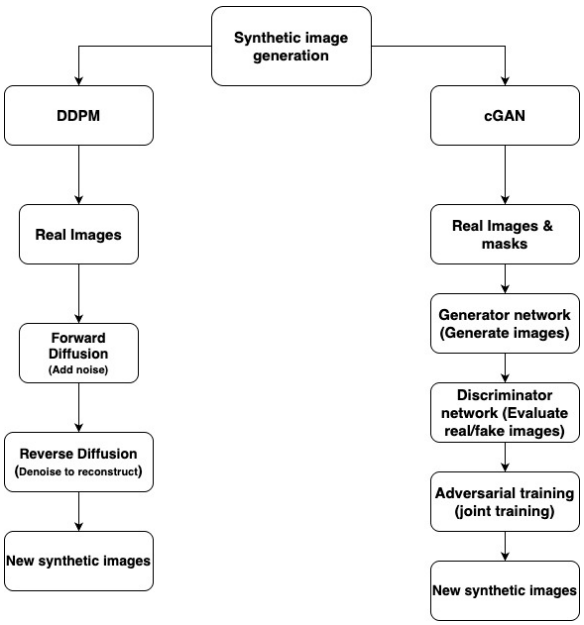

Figure 1. Generation of synthetic data

Before training the U-Net model, several pre-processing steps were implemented. To maintain consistency in input dimensions, all images were resized to 256×256. The images were normalized by dividing the pixel values by 255.0, thereby scaling the pixel values to the range of [0, 1]. All masks were converted to single-channel grayscale images so that pixel values above 127 are considered as belonging to the polyp class and values below 127 are considered as belonging to the background. Since this is known to lead to class imbalance and to enhance the model's generalizability and robustness, the image-mask pairs in the real datasets (Jha et al., 2019; Ali et al., 2023) were subject to the augmentation techniques presented in Table 1 below.

Seed synchronization was implemented so that the image and the mask undergo the same spatial transformations by resetting the random seed. This, in turn, preserves the alignment between the polyp region in the mask and its corresponding region in the image.

This augmented dataset, called pseudosynthetic data, provides a more comprehensive representation of possible variations in the input data. Finally, both images and masks are converted into tensors to match the input requirements of the deep learning framework. This step preserves the alignment and dimensions needed by the U-Net model. Below in Figure 2 are depicted the possible transformations of an image.

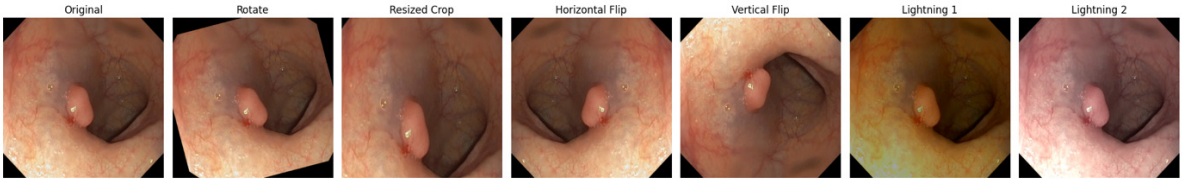

Figure 2. Generation of pseudosynthetic data – sample transformations

Table 1. Data augmentation – types of transformations

Augmentation Type	Description	Applied to	Parameters	Purpose
Spatial transformations				
Random Flips	Horizontal and vertical flips	Images & Masks	Horizontal, Vertical	Allows orientation variance
Random Rotation	Rotation within ±10°	Images & Masks	±10°	Introduces geometric diversity
Random Resized Crop	Randomly scales and crops the image	Images & Masks	Scale factor: 0.8–1.0	Varies spatial scale; enhances robustness
Color jittering				
Brightness Adjustment	Randomly adjusts brightness	Images only	±20%	Diversifies color distribution
Contrast Adjustment	Randomly adjusts contrast	Images only	±20%	Improves generalization to varying lighting conditions
Saturation Adjustment	Randomly adjusts saturation	Images only	±20%	Accounts for diverse color conditions
Hue Adjustment	Randomly shifts hue	Images only	±10%	Accounts for color variations in endoscopic imagery

2.2 Model Training and Evaluation

This study adheres to the reporting standards outlined in the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) to ensure transparency and reproducibility in the development and evaluation of the proposed polyp segmentation model. Methodological choices, including data preprocessing, model training, and performance assessment, align with the established practices in medical imaging deep learning, as recommended by CLAIM and the related guidelines (Tejani, Klontzas & Gatti, 2024; Bossuyt et al., 2015)

The data was split into training (70%), validation (15%), and test (15%) sets to ensure a thorough evaluation of the model's performance (Anon, n.d.) The split was performed by setting a specific random seed for every experiment, making it easier to compare results and reproduce experiments. A modified U-Net architecture designed for 256×256 RGB inputs was employed, which is composed of an encoder, a bottleneck layer, and a symmetric decoder with skip connections. In the encoder, each level consists of two 3×3 convolutional layers, each followed by a batch normalization layer - a modification in comparison with the original U-Net that stabilizes and accelerates training by normalizing the activations within each batch (Ronnenberger, Fischer & Brox, 2015) The model was compiled using the Adam optimizer (Kingma & Ba, 2014), with binary cross-entropy as the loss function (Ruby et al., 2020). Metrics including accuracy, precision, recall, the Dice coefficient, Intersection over Union (IoU), and the F1 score were employed to evaluate the model's performance (Anon, n.d.).

Callbacks for model checkpointing, early stopping, and Tensor Board logging were implemented to monitor and enhance the training process (Anon, n.d.). The trained model's performance was evaluated on the test set to assess its generalizability and effectiveness in segmenting unseen data.

This study places particular emphasis on the importance of precision-recall as a key metric in model evaluation, in addition to more traditional metrics. Given the clinical significance of accurately detecting polyps while minimizing false positives and false negatives, precision-recall offers a valuable insight into the model's performance, particularly in scenarios with imbalanced data, where accuracy alone may not provide a comprehensive assessment of the model's effectiveness (Erickson & Kitamura, 2021).

2.3 Qualitative Evaluation

In addition to quantitative metrics, the model's segmentation accuracy was qualitatively assessed by a team of gastroenterologists to ensure its clinical relevance. The evaluation involved three independent reviewers: a practicing gastroenterologist (the primary author), a senior gastroenterologist with over 15 years of experience, and a resident gastroenterologist in training. For this assessment, 20 randomly selected test images from the dataset were processed to generate binary polyp masks. The model's predicted outputs were thresholded to create these masks, which were then visually compared side-by-side with the corresponding ground truth masks and the original endoscopic images.

Each reviewer independently evaluated the 20 cases, focusing on the model's ability to accurately delineate polyp boundaries and distinguish polyps from surrounding mucosal tissue. The segmentation accuracy, as determined by the overlap between the predicted and ground truth masks, was consistently high, with a mean Dice coefficient of 0.91 across the evaluated samples (range: 0.89–0.94). The inter-rater agreement was assessed using Cohen's kappa coefficient, yielding a value of 0.87, which indicates a strong concordance among the reviewers. Discrepancies, observed in approximately 10% of cases, were primarily attributed to subtle differences in interpreting polyp edges in regions with a low contrast or irregular mucosal patterns. These cases were resolved through discussion, with the consensus reached on the model's performance being clinically satisfactory.

Visualizations were generated for five representative test images, each displaying the original endoscopic image, the ground truth mask, and the model's predicted mask. These visualizations, reviewed by the gastroenterologists, further confirmed the model's capability to produce precise and clinically interpretable segmentations. This qualitative evaluation

shows the model`s potential as a reliable tool for assisting gastroenterologists in polyp detection during endoscopic procedures, as it can be seen in Figures 4-7. A flowchart depicting the proposed method is displayed in Figure 3.

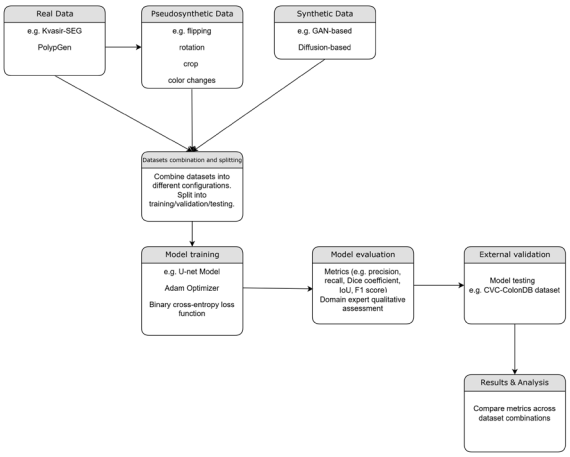

Figure 3. Flowchart of the proposed method

2.4 Experiments

All experiments were performed using Python as a programming language, Google Colab environment for implementation. The training was performed on a Nvidia A100 High-RAM GPU. After training, all the resulting models were submitted to external validation using the CVC-Colon-DB dataset comprising 612 image-mask pairs (Bernal, Sánchez & Vilarino, 2012). The dataset`s real-world applicability and diagnostic utility was validated using several performance metrics (precision, recall, the Dice coefficient, IoU, and the F1 score) and through external validation on the CVC-Colon-DB dataset (612 image-mask pairs). A gastroenterologist visually assessed the segmentation accuracy to ensure the clinical relevance of the proposed method. Pseudosynthetic data preserved traceability to the original clinical source through controlled augmentations, while synthetic data underwent visual checks for anatomical plausibility.

2.4.1 Real Data (Baseline)

The aim was to establish a baseline performance using real-world data, which would serve as a foundation for comparison with subsequent experimental setups. This experiment included a total of 4,762 real images, combining 1000 images from the Kvasir-SEG dataset and 3762 images from the PolypGen dataset.

2.4.2 Pseudosynthetic Data

The objective was to assess the model's ability to generalize from an augmented dataset reflecting a broader spectrum of conditions than the original dataset. Six augmentations were applied to each image from the combined Kvasir-SEG and PolypGen datasets, resulting in 28,572 images. Spatial transformations - flips, rotations, and resized crops - were applied to both images and masks to mimic variations in polyp appearance. Color transformations, brightness, contrast, saturation, and hue adjustment were applied only to images to simulate different lighting conditions. A custom data generator managed image loading, preprocessing (resizing and normalization), and batch creation for model training.

2.4.3 Synthetic Data (CycleGAN and Polyp-DDPM)

The aim was to assess the model's performance when trained exclusively on synthetic data, exploring whether it can supplement or replace real data when it is scarce or incomplete. The Synth-Colon dataset was used, comprising 20,017 synthetic images generated using the CycleGAN model based on the Kvasir dataset. Afterwards, one evaluated the U-Net model trained exclusively on synthetic data generated using Polyp-DDPM, a diffusion-based semantic polyp synthesis method, by training on the Kvasir dataset of 1,000 image-mask pairs for 25,000, 50,000, and 100,000 epochs. Using 5,000 masks, 20,000 new image-mask pairs were generated, which were then used to train the U-Net model.

2.4.4 Experiments Using Combinations of Datasets

a. Real and pseudosynthetic data

To assess the expected improvement from adding pseudosynthetic data to real data, the U-Net model was trained using a total of 33,334 image-mask pairs - comprising 4,762 real pairs and 28,572 pseudosynthetic pairs.

b. Real and synthetic data (cGan)

In this experiment the U-Net model was trained using the real image dataset and the cGan synthetic dataset, for a total of 24,779 pairs of images and masks.

c. Real and synthetic data (Polyp-DDPM)

Two experiments were conducted using both real and Polyp-DDPM generated datasets. In the first experiment, the real dataset was combined with synthetic data generated by a model trained for 25,000 epochs. In the second experiment, the real dataset was used along with synthetic data from a model trained for 50,000 epochs. In both experiments, the datasets consisted of 4,762 real images and 20,000 synthetic images.

d. Real, pseudosynthetic and synthetic data (cGan)

The model's performance was evaluated when it was trained on a dataset consisting of 4,762 real images, 28,572 pseudosynthetic Images and 20,017 synthetic images generated using CycleGAN.

e. Real, pseudosynthetic and synthetic data (Polyp-DDPM)

In this experiment, the Polyp-DDPM generated dataset (image generation model trained for 25000 epochs) was used along with real and pseudosynthetic datasets.

f. Real, pseudosynthetic and all synthetic data (cGan + Polyp-DDPM)

Finally, the real dataset was combined with pseudosynthetic images and synthetic data generated by both the CycleGAN and Polyp-DDPM models (the latter trained for 25,000 epochs). This resulted in a total of 73,351 image-mask pairs, which were used for training the U-Net segmentation model.

3. Results

3.1 Training on Real Data

The model was set for training for 40 epochs, with early stopping activated at epoch 24 to prevent overfitting. The final test metrics included a loss of 0.1289, an accuracy of 0.9709, a precision of 0.9003, a recall of 0.7307, a Dice coefficient of 0.7911, a IoU of 0.6593, and a F1 score of 0.7903.

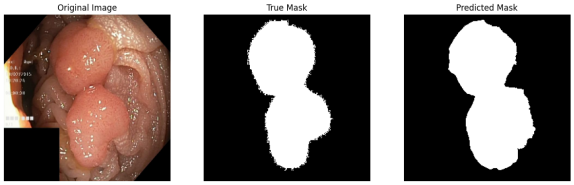

Figure 4. Qualitative assessment of segmentation accuracy – true vs predicted mask (real data)

3.2 Training on Pseudosynthetic Data

When training on pseudosynthetic data, early stopping was activated at epoch 24 during the training. The model's evaluation results show a test loss of 0.1045, a test accuracy of 0.9787, a precision of 0.9212, a recall of 0.8672, a dice coefficient of 0.8847, a IoU of 0.7948, and a F1 score of 0.8950.

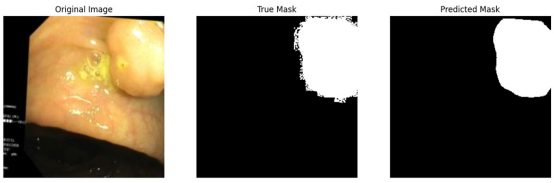

Figure 5. Qualitative assessment of segmentation accuracy – true vs. predicted mask (pseudosynthetic data)

3.3 Training on Synthetic Data (cGan)

The third experiment focused on training the model exclusively on synthetic data over 20 epochs. On the test set, the model achieved a loss of 0.0045, an accuracy of 0.9950, a precision of 0.9954, a recall of 0.9631, a Dice coefficient of 0.9809, and a IoU of 0.9625.

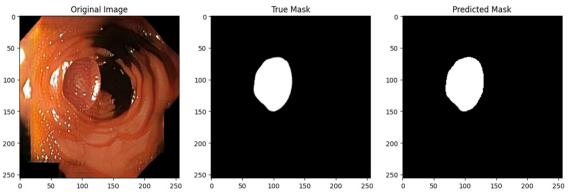

Figure 6. Qualitative assessment of segmentation accuracy – true vs. predicted mask (synthetic data – cGan).

3.4 Training on Synthetic Data (Polyp-DDPM)

A model was trained to generate synthetic images using three different settings and then the U-Net model was trained based on each generated dataset by the aforementioned generation model. The results are as follows:

Test metrics (for 25,000 epochs): a loss of 0.0259, an accuracy of 0.9915, a precision of 0.9301, a recall of 0.9218, a Dice coefficient of 0.9061, a IoU of 0.8289, and a F1 score of 0.9253.

Test metrics (for 50,000 epochs): a loss of 0.0158, an accuracy of 0.9952, a precision of 0.9675, a recall of 0.9496, a Dice coefficient of 0.9521, a IoU of 0.9087, and a F1 score of 0.9589.

Test metrics (for 100,000 epochs): a loss of 0.0112, an accuracy of 0.9962, a precision of 0.9714, a recall of 0.9633, a Dice coefficient of 0.9607, a IoU of 0.9245, and a F1 score of 0.9672.

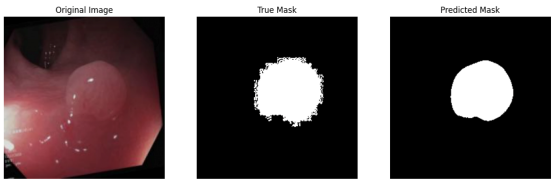

Figure 7. Qualitative assessment of segmentation accuracy – true vs. predicted mask (synthetic data – Polyp-DDPM)

3.5 Real and Pseudosynthetic Images

For real and pseudosynthetic images, the test metrics included a loss of 0.1078, an accuracy of 0.9783, a precision of 0.9349, a recall of 0.8480, a Dice coefficient of 0.8799, a IoU of 0.7875, and a F1 score of 0.8832.

3.6 Real and Synthetic Data (cGan, Polyp-DDPM)

For the experiment using real and cGan generated datasets, the test metrics are as follows: a loss of 0.0518, an accuracy of 0.9807, a precision of 0.9305, a recall of 0.7284, a Dice coefficient of 0.7312, a IoU of 0.6559, and a F1 score of 0.7745.

For real and synthetic data (Polyp-DDPM), the model was trained using two different settings:

Test metrics (for the model trained for 25,000 epochs): a loss of 0.0593, an accuracy of 0.9863, a precision of 0.9249, a recall of 0.8546, a Dice coefficient of 0.8739, a IoU of 0.7778, and a F1 score of 0.8908.

Test metrics (for the model trained for 50,000 epochs): a loss of 0.0415, an accuracy of 0.9899, a precision of 0.9503, a recall of 0.8888, a Dice coefficient of 0.9075, a IoU of 0.8326, and a F1 score of 0.9142.

3.7 Real, Pseudosynthetic, Synthetic Data

For the combination of real, pseudosynthetic, and cGan data, the values of the test metrics were as follows: a loss of 0.0589, accuracy of 0.9869, precision of 0.9509, recall of 0.8918, Dice coefficient of 0.9114, IoU of 0.8386, and F1 score of 0.9197.

For the experiment using real, pseudosynthetic, and Polyp-DDPM data, the test metrics included: a loss of 0.0776, an accuracy of 0.9811, a precision of 0.9243, a recall of 0.8450, a Dice coefficient of 0.8667, a IoU of 0.7681, and a F1 score of 0.8850.

For real, pseudosynthetic, and synthetic (cGan + Polyp-DDPM) data, test metrics included: a loss of 0.0508, an accuracy of 0.9872, a precision of 0.9389, a recall of 0.8904, a Dice coefficient of 0.9018, a IoU of 0.8228, and a F1 score of 0.9116.

The external validation data for all experiments is included in Table 2 below. A color gradient is employed to visually represent the results of the experiments carried out for different metrics.

Table 2. External validation metrics for all experiments. Color map: red indicates a lower performance, green a higher performance

Experiment	Avg. Dice Score	Avg. IoU	Precision	Recall	F1 Score
1. Real Data	0.5824	0.4951	0.8536	0.5792	0.6369
2. Pseudosynthetic Data	0.7429	0.6463	0.8911	0.6481	0.7501
3. Synthetic Data (cGan)	0.1091	0.0828	0.7246	0.0898	0.4944
4a. Synthetic Data (Polyp-DDPM, 25k epochs)	0.6226	0.5243	0.7697	0.6435	0.715
4b. Synthetic Data (Polyp-DDPM, 50k epochs)	0.5219	0.4479	0.7226	0.5119	0.7675
4c. Synthetic Data (Polyp-DDPM, 100k epochs)	0.5623	0.4802	0.8007	0.5469	0.7289
5. Real + Pseudosynthetic Images	0.7638	0.6774	0.8979	0.7535	0.7797
6. Real + Synthetic Data (cGan)	0.6548	0.576	0.8909	0.6393	0.6929
7a. Real + Synthetic Data (Polyp-DDPM, 25k epochs)	0.6593	0.5763	0.8776	0.6506	0.6882
7b. Real + Synthetic Data (Polyp-DDPM, 50k epochs)	0.6314	0.5589	0.9069	0.6133	0.6681
8. Real + Pseudosynthetic + cGan Data	0.7488	0.6695	0.8987	0.7299	0.7774
9. Real + Pseudosynthetic + Polyp-DDPM Data	0.7319	0.6465	0.8963	0.7227	0.7511
10. Real + Pseudosynthetic + cGan + Polyp-DDPM Data	0.7499	0.6687	0.8949	0.7344	0.7675

4. Discussion

Performing multiple colonoscopies on the same patient within short intervals is impractical due to ethical concerns, patient safety, and the invasive nature of the procedure. Frequent colonoscopies carry risks like bowel perforation, infection, and patient discomfort. Moreover, there's limited clinical need for short interval repeat colonoscopies, as significant pathological changes, such as polyp growth or morphological alterations, typically occur over longer periods (Rognstad et al., 2024; Zhang et al., 2021). While back-to-back colonoscopies are occasionally performed for immediate reassessment, images from successive procedures often show minimal differences, leading to a limited dataset variability. Since the polyp morphology and surrounding mucosa remain largely unchanged over short periods of time, this redundancy can hinder the development of machine learning models, which require diverse and representative datasets to perform accurately. Pseudosynthetic data, derived from the augmentation of real-world colonoscopy images, provides a solution to this challenge. By applying augmentation techniques, pseudosynthetic data introduces controlled variations to the original images while preserving the clinical characteristics of the source data. This process enhances the diversity of the dataset, simulating conditions that could occur in future colonoscopies without requiring additional invasive procedures to be performed on patients. Consequently, pseudosynthetic data allows for the creation of robust models that are trained to recognize a broader range of polyp appearances, while maintaining the traceability and clinical relevance of the original images. Synthetic data generated using GANs or diffusion models, though effective in increasing the diversity of datasets, lacks this traceability. Such models are trained on real data but produce entirely novel images that cannot be traced back to specific instances in the original dataset. This decoupling from the source data raises concerns about the interpretability and validation of the generated data, as it cannot be directly attributed to any real-world image or clinical case. Pseudosynthetic data offers a distinct advantage over purely synthetic data due to the traceability which allows a direct reference to original clinical images. Optimal traceability facilitates regulatory approval, transparency with regard to data origin, reproducibility of results, and accountability with regard to model validation. It also supports clinical adoption by improving trust in AI outputs, as the models trained with traceable data can be validated against real-world scenarios, thereby addressing key compliance and ethical concerns in AI assisted diagnostics.

The choice between pseudosynthetic and synthetic data depends largely on the designed application. If traceability, clinical relevance, and regulatory approval are priorities, then pseudosynthetic data is preferable due to its verifiable connection to real-world images. It's safer and more reliable when clinical decision-making is at stake because one can always refer back to the original patient data. On the other hand, if the goal is to build robust machine learning models that can handle a wide variety of cases, including rare or extreme examples, synthetic data can be very valuable. It allows for a broader generalization and better model training, but with the risk that some of the generated data may not be clinically relevant or reliable. There are multiple models that can be trained to detect polyps from colonoscopy images. In this paper the U-Net was analysed as it was specifically designed for medical image segmentation. Its architecture, with a contracting path for context capture and a symmetric expanding path for precise localization, makes it well-suited for tasks like polyp detection, where both the global context and local details are very important. U-Net-based models produce high-quality results even when trained on relatively small datasets, an important advantage in medical imaging, such as digestive endoscopy, where annotated data can be scarce and expensive to obtain. Its effectiveness has been widely validated across numerous studies. (Moreu, McGuinness & O'Connor, 2022; Kundu et al., 2022; Yousef et al., 2023).

In the field of medical diagnostics, synthetic data generation has become a pivotal strategy for enhancing machine learning model training, particularly when real-world data is limited. For example, Matei et al. (2023) developed a method for creating synthetic datasets using computational fluid dynamics (CFD) simulations to estimate blood pressure drops in aortic stenosis cases. By customizing a generic aortic valve model across various anatomical parameters, they generated diverse yet physiologically plausible valve shapes,

enabling the training of ML models for an accurate pressure drop estimation without relying on actual patient data. Similarly, Nawroly et al. (2024) addressed data scarcity in dysarthric speech recognition by employing a two-stage transfer learning approach that combines category-specific noise augmentation with speaker-specific data augmentation. This method effectively increased the volume and diversity of training data, leading to a notable reduction in the Word Error Rate, particularly among the severely affected dysarthric speakers. This study on colon polyp detection used diffusion-based models (Polyp-DDPM) to generate synthetic polyp images, expanding the training dataset and addressing limitations associated with data scarcity and diversity. This approach aligns with the principles demonstrated by Matei et al. (2023) and Nawroly et al. (2024) underscoring the efficacy of synthetic data in developing robust, generalizable ML models in medical diagnostics.

Overfitting is a concern in deep learning, especially when dealing with limited datasets typical in the medical imaging domain. It occurs when a model learns the details and noise in the training data to such an extent that it negatively impacts on the performance of the model on new data. This is particularly problematic for tasks like polyp detection from colonoscopy images, where the model's ability to generalize to unseen data is required for its clinical utility. Pseudosynthetic and synthetic data were used to combat overfitting by introducing additional variability that mimicked real-world conditions. This variability prevented the model from memorizing specific image patterns, which led to promoting the learning of more general features that are indicative of polyps. The implementation of early stopping mechanisms was another way to limit overfitting. By monitoring the validation loss and halting the training process when no improvement was observed for 10 epochs, the model was protected from the risk of becoming overly attuned to the training data and reached the right balance between learning from the data and maintaining the ability to perform well on new, unseen data.

The model trained on pseudosynthetic data showed good generalization capabilities and was able to capture the essential features of the polyps effectively, even on unseen data. Training solely on CycleGAN-generated synthetic data leads to high test performance, but the poor generalization to external validation data indicates a severe overfitting issue. The model apparently learns artifacts specific to synthetic images and struggles to translate this knowledge into real-world data. The diffusion-based synthetic data led to a better generalization capability than CycleGAN, but it still underperforms with regard to external validation in comparison with real or augmented data. This suggests that diffusion models create higher-quality synthetic images, but further tuning might be needed to ensure that the synthetic data mimics real-world variability more accurately. The inclusion of synthetic data allows for a significant improvement in both training and generalization in comparison with baseline data, whereas combining real and pseudosynthetic data leads to better results than the ones obtained in the other training experiments. Synthetic data helps to enhance model performance, but it cannot replace real data alone. While synthetic data is useful for improving model training results, care must be taken to avoid overfitting to synthetic features, as it was the case with the external validation results for the CycleGAN-only training. Further research should explore how to refine synthetic data generation techniques to better reflect real-world variability.

Mode collapse is a common issue in Generative Adversarial Networks (GANs) where the generator, instead of producing diverse outputs, repeatedly generates limited variations of a few specific outputs. This occurs because the generator learns to focus on producing examples that repeatedly fool the discriminator, but these examples represent only a small portion of the overall data distribution. As a result, the generator fails to capture the full diversity of the target data, leading to outputs that lack variety. This problem can be particularly detrimental for tasks where diversity is essential, such as image generation or data synthesis. When mode collapse happens, the model essentially "cheats" by sticking to a few types of data points that consistently trick the discriminator, rather than learning the complete distribution of the dataset. Several techniques, such as improving the training dynamics between the generator and discriminator, by using regularization, or employing different

loss functions, have been proposed to mitigate mode collapse, but it remains a challenge in GAN development (Su et al., 2021). Based on the analysis carried out, CycleGAN-generated data faced challenges related to mode collapse, producing limited variations that reduced dataset diversity and led to overfitting. This lack of variability caused the model to learn synthetic artifacts and not clinically relevant features, which weakened its generalization ability for external datasets. In contrast, diffusion-based models generated more diverse and anatomically realistic samples and better mimicked real-world variability. These improvements enhanced model robustness and supported a more reliable performance in external validation. The recent progress in diffusion-based models has overcome the mode collapse issue, producing diverse, high-quality images that outperform GANs. However, despite its effectiveness in generating varied images, this method incurs high computational costs for training and inference (Durall et al., 2021). Diffusion based data generation required approximately 2 times more computational resources (determined time-wise) than GANs due to the iterative denoising steps involved in sample generation. To scale this method in resource-constrained environments, it is recommendable to make use of pre-trained diffusion models for transfer learning, a reduced resolution training, and distributed computing frameworks. These approaches can lower the computational costs while preserving model performance. The external validation dataset CVC-Colon-DB was selected due to its widespread use in polyp segmentation research and its inclusion of diverse polyp shapes, sizes, and textures. However, it may not fully capture the heterogeneity observed in clinical settings, particularly variations related to patient demographics, endoscopy equipment, or imaging protocols. This limitation highlights the need for future studies to validate models using multicenter datasets to ensure a broader generalizability and clinical applicability. In this paper a modified U-Net architecture was employed, which was chosen for its proven efficacy in medical image segmentation. Although suitable as a baseline model, other architectures (e.g. transformers or attention-based models – attention-based U-Nets) might achieve a better performance or show different responses to synthetic and pseudosynthetic data. Comparing multiple architectures would be a valuable step toward more complex conclusions, which are architecture-neutral.

Moreover, in recent years, several studies have explored the use of synthetic data to enhance polyp detection and segmentation in colonoscopy images. For instance, Shin et al. (2018) used conditional adversarial networks to generate synthetic polyp images from normal colonoscopy images, improving polyp detection performance by augmenting the training datasets with realistic polyp appearances. Similarly, Haugland et al. (2023) utilized a CycleGAN-based framework to translate white-light imaging (WLI) to synthetic narrow-band imaging (SNBI), enhancing polyp detection by leveraging the improved visibility of polyps in NBI. These approaches highlight the potential of synthetic data in addressing data scarcity and improving model performance in medical image analysis.

Future research should focus on integrating additional capabilities into U-Net models, such as real-time pathology detection when incorporated into endoscopy systems. This would enable endoscopists to receive immediate feedback or alerts for the polyps detected during live procedures, significantly enhancing the clinical workflow. Beyond polyp detection, there is a substantial opportunity to extend the application of AI models to post-detection tasks, including polyp classification (e.g as benign vs. malignant), size estimation, and even recommending optimal treatment pathways based on the detected abnormalities.

5. Conclusion

This study is among the first to systematically evaluate the impact of synthetic data and pseudosynthetic data—a term referring to data that simulates image variability as if obtained from multiple endoscopies—on enhancing the diagnostic accuracy of deep learning models for colon polyp detection. This paper investigated and confirmed the potential of pseudosynthetic and synthetic data as effective tools for addressing the scarcity and lack of diversity typical of real-world datasets, for enhancing model generalization through data augmentation, and addressing ethical

issues related to patient privacy in AI-assisted diagnostic environments. The experiments carried out demonstrated that the U-Net model performs better when trained on synthetic and pseudosynthetic data than when trained solely on real data, highlighting the importance of extensive and diverse training datasets in the field of digestive endoscopy. Notably, models trained exclusively on pseudosynthetic data outperformed those trained on a mix of synthetic and pseudosynthetic sources.

Furthermore, in alignment with the findings from other studies, the models trained on synthetic data generated using diffusion algorithms showed a superior performance in comparison with those trained on data produced by GANs.

REFERENCES

Ahmad, O.F., Soares, A.S., Mazomenos, E. et al. (2019) Artificial intelligence and computer-aided diagnosis in colonoscopy: Current evidence and future directions. *The Lancet Gastroenterology and Hepatology*. 4(1), 71-80. https://doi.org10.1016/S2468-1253(18)30282-6.

Ali, S., Jha, D., Ghatwary, N. et al. (2023) A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*. 10, art. no. 75. https://doi.org/10.1038/s41597-023-01981-y.

Anon. (n.d.) *Keras 3 API Documentation*. https://keras.io/api/ [Accessed 1st September 2024].

Barua, I., Vinsard, D.G., Jodal, H.C. et al. (2021) Artificial intelligence for polyp detection during colonoscopy: a systematic review and meta-analysis. *Endoscopy*. 53(3), 277-284. https://doi.org/10.1055/a-1201-7165.

Bernal, J., Sánchez, J. & Vilarino, F. (2012) Towards Automatic Polyp Detection with a Polyp Appearance Model. *Pattern Recognition*. 45(9), 3166-3182. https://doi.org/10.1016/j.patcog.2012.03.002.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E. et al. (2015) STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *Radiology*. 277(3), 826-832. https://doi.org/10.1148/radiol.2015151516.

Chen, M.Z.; Devan Nair, H.; Saboo, A. et al. (2022) A single-centre audit: Repeat pre-operative colonoscopy. *ANZ Journal of Surgery*. 92(10), 2571-2576. https://doi.org/10.1111/ans.17813.

Dorjsembe, Z., Pao, H.K. & Xiao, F. (2024) Polyp-DDPM: Diffusion-Based Semantic Polyp Synthesis for Enhanced Segmentation. In: *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 15-19 July 2024, Orlando, USA*. New York, USA, IEEE. https://doi.org/10.1109/EMBC53108.2024.10782077.

Durall, R., Chatzimichailidis, A., Labus, P. et al. (2021) Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues. In: *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISGRAPP 2021), 8-10 February 2021, Online*. Vol. 4. Setubal, Portugal, SciTePress. pp. 211-218.

Erickson, B.J. & Kitamura, F. (2021) Magician's Corner: Performance metrics for Machine Learning Models. *Radiology: Artificial Intelligence*. 3(3), art. no. e200126. https://doi.org/10.1148/ryai.2021200126.

Haugland, M. R., Qadir, H. A. & Balasingham, I. (2023) Deep learning for improved polyp detection from synthetic narrow-band imaging. In Iftekharuddin, K. M. & Chen, W. (eds.) *Medical Imaging 2023: Computer-Aided Diagnosis* (*SPIE Medical Imaging, 19-24 February 2023, San Diego, California, United States*), vol. 12465. Bellingham, Washington, USA, SPIE. https://doi.org/10.1117/12.2653048.

Herszényi, L. (2019) The "Difficult" Colorectal Polyps and Adenomas: Practical Aspects. *Digestive Diseases*. 37(5), 394-399. https://doi.org/10.1159/000495694.

Ho, J., Jain, A. & Abbeel, P. (2020) Denoising Diffusion Probabilistic Models. In: *NIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems, December 6 - 12, 2020, Vancouver, Canada.* Red Hook, NY, USA, Curran Associates Inc.. pp. 6840-6851.

Isola, P., Zhu, J.-Y., Zhou, T. et al. (2017) Image-to-Image Translation with Conditional Adversarial Networks. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 21-26 July 2017, Honolulu, HI, USA.* New York, USA, IEEE.

Jha, D., Smedsrud, P.H., Riegler, M.A. et al. (2019) Kvasir-SEG: A Segmented Polyp Dataset. In: Ro, Y. M., Cheng, W.-H., Kim, J., Chu, W.-T., Cui, P., Choi, J.-W., Hu, M.-C. & De Neve, W. (eds.) *MultiMedia Modeling* (*26th International Conference, MMM 2020, 5-8 January 2020, Daejeon, South Korea, Part II*, part of the book series *Lecture Notes in Computer Science (LNCS)*, vol. 11962). Cham, Switzerland, Springer International Publishing, pp. 451-462.

Jiang, W., Xin, L., Zhu, S. et al. (2023) Risk Factors Related to Polyp Miss Rate of Short-Term Repeated Colonoscopy. *Digestive Diseases and Sciences*. 68, 2040-2049. https://doi.org/10.1007/s10620-023-07848-x.

Kingma, D.P. & Ba, J. (2014) Adam: A Method for Stochastic Optimization. To be published in: *ICLR 2015*. [Preprint] https://arXiv:1412.6980 [Accessed 1st September 2024].

Kundu, S., Karale, V., Ghorai, G. et al. (2022) Nested U-Net for Segmentation of Red Lesions in Retinal Fundus Images and Sub-image Classification for Removal of False Positives. *Journal of Digital Imaging*. 35(5), 1111-1119. https://doi.org/10.1007/s10278-022-00629-4.

Matei, T.I., Popescu, A.B., Nita, C.I. et al. (2023) CFD-based Synthetic Data Generation for Machine Learning based Pressure Drop Assessment in Aortic Stenosis. *Studies in Informatics and Control*. 32(4), 49–58, https://doi.org/10.24846/v32i4y202305.

Moreu, E., McGuinness, K. & O'Connor, N. E. (2022) *Synthetic data for unsupervised polyp segmentation.* arXiv e-prints. https://arxiv.org/abs/2202.08680

Morgan, E., Arnold, M., Gini, A. et al. (2023) Global burden of colorectal cancer in 2020 and 2040: Incidence and mortality estimates from GLOBOCAN. *Gut*. 72(2), 338-344. https://doi.org/10.1136/gutjnl-2022-327736.

Morrow, L. & Greenwald, B. (2022) Healthy Food Choices, Physical Activity, and Screening Reduce the Risk of Colorectal Cancer. *Gastroenterology Nursing*. 45(2), 113-119. https://doi.org/10.1097/SGA.00000000000006.15

Nawroly, S.S., Popescu, D. & Thekekara Antony, M.C. (2024) Category-Based and Target-Based Data Augmentation for Dysarthric Speech Recognition Using Transfer Learning. *Studies in Informatics and Control*. 33, 83–93. https://doi.org/10.24846/v33i4y202408.

Ronneberger, O., Fischer, P.& Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (*18th International Conference, 5-9 October 2015, Munich, Germany, Proceedings, Part III*), part of the LNCS series, vol. 9351). Cham, Switzerland, Springer, pp. 234, 241.

Rognstad, Ø.B., Botteri, E., Hoff, G. et al. (2024) Adverse events after colonoscopy in a randomized colorectal cancer screening trial. *BMJ Open Gastroenterology*. 11(1), art. no. e001471.

Ruby, U., Theerthagiri, P., Jacob, J. et al. (2020) Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering.* 9(4), 5393-5397. https://doi.org/10.30534/ijatcse/2020/175942020.

Shao, L., Yan, X., Liu, C. et al. (2022) Effects of AI-assisted colonoscopy on adenoma miss rate/adenoma detection rate: A protocol for systematic review and meta-analysis. *Medicine (Baltimore)*. 101(46), art. no. e31945. https://doi.org/10.1097/MD.0000000000031945.

Shin, Y., Qadir, H.A. & Balasingham, I. (2018) Abnormal Colon Polyp Image Synthesis Using Conditional Adversarial Networks for Improved Detection Performance. *IEEE Access*. 6, 56007-56017. https://doi.org/10.1109/ACCESS.2018.2872717.

Su, R., Zhang, D., Liu, J. et al. (2021) MSU-Net: Multi-Scale U-Net for 2D Medical Image Segmentation. *Frontiers in Genetics*. 12, art. no. 639930. https://doi.org/10.3389/fgene.2021.639930.

Sullivan, B.A., Noujaim, M. & Roper, J. (2022) Cause, Epidemiology, and Histology of Polyps and Pathways to Colorectal Cancer. *Gastrointestinal Endoscopy Clinics of North America*. 32(2), 177-194. https://doi.org/10.1016/j.giec.2021.12.001.

Tejani, A.S., Klontzas, M. E. & Gatti, A. A. (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 Update. *Radiology: Artificial Intelligence.* 6(4), art. no. e240300. https://doi.org/10.1148/ryai.240300.

van Liere, E., Jacobs, I.L.; Dekker, E. et al. (2023) Colonoscopy surveillance in Lynch syndrome is burdensome and frequently delayed. *Familial Cancer*. 22(4), 403-411. https://doi.org/10.1007/s10689-023-00333-4.

Wilhelmi, M., Burkhart, A. & Netzer, P. (2021) Kolonkarzinom - Wie können wir die Prävention verbessern? [Colorectal carcinoma - How can we improve prevention?]. *Therapeutische Umschau*. 78(2), 61-72. https://doi.org/10.1024/0040-5930/a001240.

Williamson, S.M. & Prybutok, V. (2024) Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare. *Applied Sciences*. 14(2), art. no. 675. https://doi.org/10.3390/app14020675.

Yousef, R., Khan, S., Gupta, G. et al. (2023) U-Net-Based Models towards Optimal MR Brain Image Segmentation. *Diagnostics* (*Basel*). 13(9), art. no. 1624. https://doi.org/10.3390/diagnostics13091624.

Zhang, Q., Shen, Y., Xu, J. et al. (2021) Clear colonoscopy as a surveillance tool in the prediction and reduction of advanced neoplasms: A randomized controlled trial. *Surgical Endoscopy*. 35(8), 4501-4510. https://doi.org/10.1007/s00464-020-07964-z.

Zhu, J.-Y., Park, T., Isola, P. et al. (2017) Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In: *2017 IEEE International Conference on Computer Vision* (*ICCV*)*, 22-29 October 2017, Venice, Italy*. New York, USA, IEEE. https://doi.org/10.1109/ICCV.2017.244.