Enhanced Model for Targeted False Positive Synthesis via Detector-guided Adversarial Diffusion Attacker for Robust Polyp Detection

Dalal Aldowaihi
Student ID: g202514270
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Many polyp detection models face a critical struggle with high false-positive rates due to the limited scale and diversity of existing datasets. These false alarms can distract clinicians and slow down early detection of colorectal cancer. To address this limitation, the DADA framework was proposed to improve adversarial diffusion-based generation of synthetic polyp images. In this study, the DADA framework was enhanced by introducing three key improvements. First, the perturbation factor α was adapted accordito guide the model during the generation process more effectivelyturbations in areas that are commonly associated with false positito anatomical regions, allowing the model to apply stronger perturbations in areas the model during the generation process. Third, a OneCycleLR learning-rate scheduler was integrated to stabilize training and improve convergence. These enhancements were evaluated on the Kvasir dataset and ETIS-Larib Polyp DB. Across experiments, the improved framework consistently boosts detector performance, most notably in F1-score, recall, and precision, demonstrating clear gains over the baseline model.**

***Index Terms*—Neural Networks, Deep Learning, Data synthesis, Adversarial diffusion, Model Enhancement, Colorectal Polyp Detection, Medical Imaging, Adversarial Attack, Polyp Detection, Segmentation Masks**

I. Introduction

A. Background and Significance

Colorectal cancer is one of the most difficult and life-threatening cancers to diagnose, often starting as polyps that form along the inner lining of the colon [1] [2]. Detecting these polyps early, before they progress into malignant tumours, is critical to reducing colorectal cancer related mortality [1] [3]. Despite colonoscopy remaining the clinical standard, it relies heavily on clinicians manually inspecting the video feed from a camera equipped endoscope [2]. The standard examination is time consuming, labour intensive and vulnerable to human mistake due to differences in experience level and clinicians' skills as well as the wide variation in polyp shape, size, appearance and the inherently complex anatomy of the colon [1] [2]. These obstacles highlight the demand for robust reliable automated polyp detection systems to support accurate diagnosis in the early stages [3].

Recent advances in machine learning research have led to remarkable developments in automated polyp detection systems, driven by powerful deep learning models designed to decrease false positives and ease clinicians' workload [2] [3]. In medical image analysis, various architectures such as YOLO-based [4], Transformer-based [5], and R-CNN-based [6] [3] models have demonstrated strong performance.

However, the colon contains folds, specular highlights, shadows and circular structures that closely resemble polyps, as well as the natural similarity between polyps and surrounding tissue, making accurate detection challenging for automated detection systems, which recent developments in detection models still face considerable challenges due to the visual complexity [1]. In addition, detection performance is constrained by the high variability of polyp appearance and the colon environment's dynamic nature [3]. Several studies have been conducted to mitigate problems such as reflection artifacts [1] and polyp size variability, development remains constrained by the absence of sufficiently diverse and comprehensive datasets [3].

To address these limitations, the Detector-guided Adversarial Diffusion Attacker (DADA) framework was introduced by [3]. DADA is a novel framework designed to generate synthetic images that are not real polyps but highly realistic "polyp-like" negative samples specifically designed to confuse the detector and reduce false-positive predictions, thereby improving polyp detection models. Comprehensive evaluations of DADA compared with the current state-of-the-art on in-house and Kvasir datasets confirm the method's superiority, with the synthesized samples boosting detector F1-scores by at least 2.6% and 2.7% respectively over the baselines [3].

To enhance the DADA framework, this research will refine the DADA framework's loss function using the region-adaptive DADA loss function, introduce OneCycleLR learning rate schedulers to stabilize training and promote smoother convergence and introduce region-specific perturbation strength, where the perturbation factor α is adaptively adjusted according to anatomical regions. In this approach, stronger perturbations are applied to areas that have historically produced higher false-positive rates, enabling the model to learn more robust discrimination in the challenging regions and ultimately reduce false-positive predictions.

B. Problem Statement

Despite significant progress in polyp detection, existing models still frequently produce false positives. Although generative models have shown promise by creating diverse positive samples and alleviating data limitations, they continue to struggle with false-positive errors. During a typical 15 - 20 minute colonoscopy, true polyp detections are rare compared to the large number of normal frames, while current computer-aided systems can generate up to five false alarms per minute [7]. A high false-positive rate leads to a "crying wolf effect" where clinicians are distracted and desensitised to the alerts [3]. These alarm not only increases clinicians load as well as raise diagnostic mistakes and the risk of unnecessary interventions [3] [8].

Additionally, every patient's anatomy is different and the quality of bowel preparation can differ significantly. These factors create a complex and unpredictable visual environment inside the colon, making it even more challenging for detection systems to avoid false positives [3].

To address these challenging, Zhou et al. [3] proposed the Detector-guided Adversarial Diffusion Attacker (DADA) framework for targeted false positive, which enables the generation of synthetic images that are not real polyps but highly realistic "polyp-like" negative samples specifically designed to confuse the detector and reduce false-positive predictions. DADA achieves this by combining a diffusion model for image synthesis with an adversarial attack mechanism that forces the generated sample to appear convincingly polyp-like to the detector. Nevertheless, the effectiveness and stability of this approach can still be further improved.

C. Objectives

The primary objectives of this research are to enhance the DADA framework for polyp detection models by reducing false-positive rates and improving dataset diversity. To achieve this, the following modifications will be introduced:

1) Refine the region-adaptive DADA loss function
2) Adapt the perturbation factor α across anatomical regions, applying stronger perturbations in areas prone to false positives.
3) Apply OneCycleLR Scheduler for stable training and improve convergence.
4) Evaluate the performance of the proposed enhancements through experiments on Kvasir and ETIS-Larib Polyp DB dataset.

D. Scope of Study

This study focuses on enhancing the DADA framework that generate challenging negative training samples to reduce false positives rate. The evaluation is conducted using the Kavsir and ETIS-Larib Polyp DB dataset, as the polyps detection performance is often limited by the datasets size and diversity . A background-only denoiser is developed that extracts negative patterns from common polyp datasets, coupled with a DADA module that adversarially guides the diffusion process to synthesise realistic and challenging false-positive samples.

II. Literature Review

Generative adversarial networks (GANs) have been widely explored to address the lack of diverse polyp detection datasets [9]. Early GAN-based approaches, such as conditional GANs [9], modified pix2pix [11], and ControlPolypNet [11], focused on synthesizing realistic polyp images to expand training sets and improve both detection and segmentation performance [3]. For instance, DS-GAN has shown success in producing realistic small objects, helping mitigate the challenging of detecting small polyps [10].

More recently, diffusion probabilistic models (DDPMs) have emerged as a powerful alternative to GANs, benefiting from stable training and the ability to generate high-quality, diverse samples [13]. DDPMs models learn to reverse a gradual noise-adding diffusion process using a parameterised Markov chain trained with variational inference, enabling them to generate high-quality samples through a simple Gaussian-based neural network formulation [13]. Background-Only Diffusion Model (BG-De)is a modified DDPM trained to learn only background patterns from colonoscopy images, where all polyp regions are masked out during training, ensuring the model does not accidentally learn true polyp shapes [3]. This forces the model to focus on learning diverse non-polyp background patterns, ensuring that no polyp related information leaks into the generation of polyp-like distractions [3].

On the detection side, YOLO remains one of the most widely adopted models due to its high speed and efficiency. Introduced by Redmon et al. [11], YOLO reframed object detection as a direct regression problem, predicting bounding boxes and class probabilities directly from pixel data. Its simplicity and fast inference have led many researchers to adapt YOLO variants for polyp detection. For example, Guo et al. [14] reduced false-positive rates in YOLOv3 through iterative retraining, while Cao et al. [15] and Pacal et al. [16] incorporated feature fusion, transfer learning, and advanced loss functions to improve small-polyp detection. Other works extended YOLOv4 and YOLOv5 through multi-scale architectures [17], attention mechanisms, and mosaic augmentation [18], achieving accuracy comparable to or exceeding Faster R-CNN. Additional research has leveraged GAN-generated data, hyperparameter optimization, and quantization techniques to further enhance YOLO's performance on resource-limited hardware. Overall, YOLO-based models continue to achieve strong and reliable performance in polyp detection [2].

Beyond YOLO, transformer-based architectures have also gained prominence. DETR [19] introduces a fundamentally different perspective by treating object detection as a direct set prediction problem. By eliminating components such as anchor boxes and non-maximum suppression manually, and instead using a transformer encoder–decoder with a bipartite matching loss, DETR models relationships between objects and the full scene, producing all detections in parallel [3].

However, generative methods focused on synthesizing positive samples, offering limited solutions for reducing false positives. The detector models continue to produce several false alarms

per minute. Such as high false-positive rates increase clinician workload and raise the risk of unnecessary interventions, further challenges compounded by patient variability and the inherently complex visual structure of the colon [7], [8].

III. Proposed Methodology

This section describes the proposed improvements to enhance the existing model. It outlines the challenges and limitations of the current approach, the key enhancements made to address these challenges, data processing techniques, as well as the loss functions and optimization strategy used to improve performance.

A. Existing Model and Challenges

The baseline model used in this work is the DADA framework [3], which was proposed to improve polyp detection by generating high value false positive images. Although effective, its ability to produce targeted distractors can be further strengthened. As shown in Fig 1, DADA combines diffusion models with adversarial guidance to synthesize realistic negative samples that intentionally mislead a pretrained detector. The framework consists of three main components:

- Background-Only Diffusion Model (BG-De):
  BG-De is a modified DDPM trained only on background regions. Polyp areas are masked out, ensuring that the model learns pure background textures without encoding polyp morphology.
- Adversarial Guidance (DADA Module):
  During sampling, adversarial gradients from a pretrained detector for instance YOLO or DETR are injected into the diffusion process. These gradients guide the model toward generating features that the detector is likely to misclassify as polyps, producing high-value false-positive samples.
- Local Inpainting for Realistic Integration:
  Only a chosen region of the original image is modified through inpainting, while the rest remains unchanged. This preserves anatomical consistency and ensures the synthetic false positives appear realistic.

The DADA framework combines various loss functions, such as detection, classification and localization loss, which can lead to oscillating training curves and difficulty in convergence. This shortage of smooth optimization elevates the demand for a technique that can balance learning effectively. These limitations highlight the need for further refinement.

B. Proposed Enhancements

To adress the limitation in the original DADA framework [3], three methodological and practical enhancements were introduced aimed at boosting the diversity of negative samples and enhancing training stability in order to help the model to detect challenging cases that will improve the effectiveness and robustness of the DADA framework for reducing false positives in polyp detection. The main contributions summarized as follows.

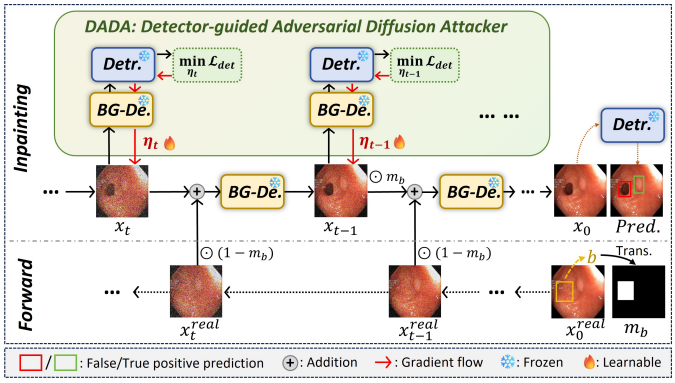

Fig. 1. The inference pipeline consists of three main components: the background-only denoiser (BG-De), a pretrained polyp detector (Detr), and the detector-guided adversarial diffusion attacker (DADA). A selected region of the real image is inpainted to generate new negative samples. BG-De reconstructs only the background, while DADA nudges the generation toward patterns that the detector is likely to mistake for polyps. For simplicity, the training processes of BG-De and the detector are not shown in this figure. [3].

Firstly, the original DADA loss function was refined and replace it with a more effective formulation "Region-adaptive DADA loss function" that strengthens the adversarial signal during diffusion sampling. This improved loss helps the model guide the diffusion process more reliably toward high-value false-positive patterns while preserving anatomical realism. By improving gradient flow especially in challenging anatomical regions the model generates synthetic samples that are more realistic and more likely to confuse the detector, making them valuable for hard negative mining.

Secondly, we introduce an anatomically adaptive perturbation strategy that adjusts the perturbation factor α based on the underlying anatomical characteristics of the image. Regions that commonly trigger false positives, such as folds, shadows, and specular highlights, receive stronger perturbations, while more uniform background areas receive weaker ones to avoid over distortion. This adaptive technique construct synthetic false-positive samples that realistic and clinically meaningful which improve the detector's ability to recognize misleading patterns during training.

Lastly, to further stabilize optimization, the OneCycleLR learning-rate scheduler was incorporated, which is particularly effective for diffusion models that are computationally sensitive and prone to oscillations during adversarial sampling. OneCycleLR begins with a low learning rate, gradually increases it to a peak, and then anneals it back toward zero over the course of the training run to encourage faster early convergence, improve generalization, and prevent instability such as exploding gradients, especially when paired with region-adaptive perturbations.

These enhancements collectively can help generate diverse and challenging negatives to confuse the detector in order to improve detector robustness.

C. Loss Function and Optimization

The DADA framework uses two main loss functions Background-Only Diffusion Loss (BG-De Loss) and Detector-Guided Adversarial Loss (DADA Loss). Background-Only Diffusion Loss to ensure that the diffusion model learns only the background distribution, the loss is computed exclusively on non-polyp regions using the ground-truth bounding-box mask m_{gtb}, and $(1 - m_{gtb})$ masks out the polyp region so that only background pixels contribute to training. Detector-Guided Adversarial Loss, during adversarial sampling, DADA treats a user-defined region b as an "illusory" ground-truth polyp. The adversarial loss is computed using the frozen detector, combining both classification and localization loss. In addition, adversarial perturbation update where DADA injects adversarial gradients into each denoising step by updating a perturbation variable η_t with a PGD-like rule where α is the perturbation step size. Adversarial Denoising Step with adversarial perturbation applied, the denoising step is 1000. The proposed framework combines multiple losses into a unified multitask learning setup. The total loss function comprises three key components:
Standard detection losses with two DADA-specific losses were combined, which include anatomical consistency and false-positive generation, while dynamically adjusting the perturbation strength α based on anatomical region type. The loss weights evolve throughout training to prioritise realism early on and stronger adversarial false-positive generation later. The enhanced "Region-Adaptive DADA Loss" for Adversarial Diffusion:
.....

Region-Adaptive DADA loss is redesigned loss function that adjusts its behavior based on the anatomical characteristics of the colonoscopy region being modified to improve DADA's ability to generate high-value false positives while keeping the synthetic image anatomically realistic. Instead of treating all regions equally, it adjust the strength perturbation since different areas (such as folds, lumen edges, specular highlights, or flat mucosa) produce false positives at different rates, so the loss function applies stronger or weaker perturbation weight depending on the region. For instance, folds and shadows often needs stronger perturbation because they fool detectors, while flat mucosa needs weaker perturbation as it is visually stable. This helps the diffusion model generate more realistic and clinically meaningful false-positive samples, while avoiding distortions in stable background regions. It also ensures that perturbations remain anatomically consistent and target locations that are most likely to confuse the detector. For an anatomical region $r \in \{\text{folds}, \text{lumen}, \text{specular}, \text{mucosa}\}$, the perturbation strength is defined as:

$$\alpha(r) = \begin{cases} \alpha_{\text{fold}} & r = \text{colon folds}, \\ \alpha_{\text{lumen}} & r = \text{lumen edge}, \\ \alpha_{\text{spec}} & r = \text{specular highlight}, \\ \alpha_{\text{mucosa}} & r = \text{flat mucosa}. \end{cases}$$

The adaptive weights follow the ordering:

$$\alpha_{\text{fold}} > \alpha_{\text{spec}} > \alpha_{\text{lumen}} > \alpha_{\text{mucosa}},$$

as colon folds typically produce the highest number of false positives and therefore require stronger perturbations.
Region-Adaptive DADA Loss is defined as:

$$\mathcal{L}_{\text{RA-DADA}} = \mathcal{L}_{\text{det}} + \lambda_{\text{anat}}(e)\, \mathcal{L}_{\text{anat}} + \lambda_{\text{fp}}(e)\, \mathcal{L}_{\text{fp}},$$

where the standard detector loss is:

$$\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{cls}}.$$

This is the loss used in YOLO to ensure bounding box accuracy, correct objectness score and correct classification of polyp versus background. Here, $\lambda_{\text{anat}}(e)$ and $\lambda_{\text{fp}}(e)$ are epoch-dependent weights that gradually adjust the contributions of the anatomical-consistency loss $\mathcal{L}_{\text{anat}}$ and the false-positive generation loss $\mathcal{L}_{\text{fp}}$.
Anatomical Consistency Loss $\lambda_{\text{anat}}(e)$ is combine the perceptual loss and style loss to ensure the synthetic region still looks like real colon tissue, even after adversarial perturbation. So, the synthetic patch remains realistic, clinically plausible and consistent with colon anatomy. The weight decreases over time because early training needs stronger shape/style constraints, while later epochs allow more freedom for generating tricky false positives.
False-Positive Generation Loss $\lambda_{\text{fp}}(e)$ pushes the detector to misclassify the perturbed region as a polyp. It amplifies features that the detector mistakenly responds to, such as round textures, bright boundaries, polyp-like shadows and distorted tissue reflections. The weight increases over time, making the attack stronger as the model becomes more stable.
However, this loss function used with OneCycleLR scheduler, the OneCycleLR is a learning rate scheduler that starts with a low learning rate, then intentionally increases the learning rate to a maximum value, then decreases it back to a very small value that will adjust momentum in the opposite direction. Because the diffusion models are very sensitive to learning rate, OneCycleLR help BG-De diffusion training stay stable, prevents collapse, prevents exploding gradients in early epochs, improves image quality, and ensures accurate denoising across all diffusion time steps, which directly strengthens DADA's ability to generate realistic false-positive samples. This "cycle" happens within one full training run, not multiple epochs, which helps to stabilize adversarial diffusion training, improve convergence when using region-adaptive losses and speed up training when the training cannot afford 200k–300k diffusion steps

IV. Experimental Design and Evaluation

This section outlines the datasets, evaluation metrics, and training configuration, followed by the experimental results. It also includes baseline comparisons and an ablation study to evaluate the contribution of each model component.

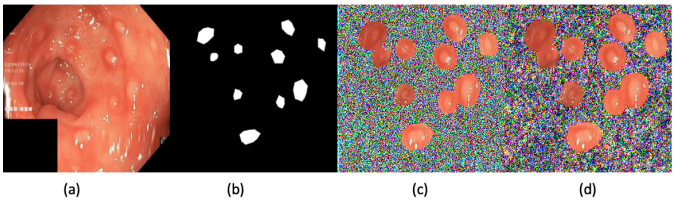

Fig. 2. Preprocessing workflow: (a) original image, (b) polyp mask, (c–d) polyp patches over Gaussian noise representing forward diffusion initialization used for DADA-based synthetic false-positive generation.

A. Datasets and Preprocessing

This study performs experiments on two datasets, as follows:

- The Kvasir [20] dataset is a public medical imaging dataset that contains 1,000 polyp images, each image includes a segmentation mask. Since DADA needs bounding boxes, the authors derived them by converting the segmentation masks into ground-truth bounding boxes around each polyp.
- The ETIS-Larib Polyp DB [21] dataset is a set of 196 high-resolution colonoscopy images containing polyps, which include challenging clinical cases with diverse imaging conditions, including specular highlights and complex backgrounds, and ground truth boxes have been manually labelled by expert endoscopists.

Several preprocessing steps are applied, all images are resized to 256×256 resolution to standardize the input dimensions for the background-only denoiser (BG-De) and the detection networks. For ETIS-Larib, mage size adjusted to 512×512 to preserve high-resolution details. For each annotated bounding box, a binary mask m_b wad constructed defined as:

$$m_b(i,j) = \begin{cases} 1, & \text{if } (i,j) \text{ is inside the bounding box,} \\ 0, & \text{if } (i,j) \text{ is outside the bounding box.} \end{cases}$$

which is essential for BG-De training (background-only learning) and DADA inpainting (local modification only). Finally, during BG-De training, polyp regions are masked via $(1 - m_{gtb})$ to ensure that only background pixels contribute to the diffusion reconstruction loss. This prevents the model from learning polyp structures and enforces learns only background textures. Kvasir and ETIS-Larib datasets are randomly split into training, validation, and testing following an 8:1:1 ratio.

B. Performance Metrics

The paper evaluates the detection performance using three standard metrics in object detection: Precision, Recall and F1-Score.Compared with the four state-of-the-art methods in Adversarial attack and diffusion-based image synthesis are APGD, FAB, RePaint and LaMa, APGD (Auto-Projected Gradient Descent) and FAB (Fast Adaptive Boundary attack) are adversarial attack methods, whereas RePaint and LaMa are image inpainting methods. Additional metrics were introduced, such as False Positive Generation Rate (FPGR) and Fréchet Inception Distance (FID), FPGR measures the proportion of generated images causing false positives, while FID quantifies the distribution discrepancy between real and generated images

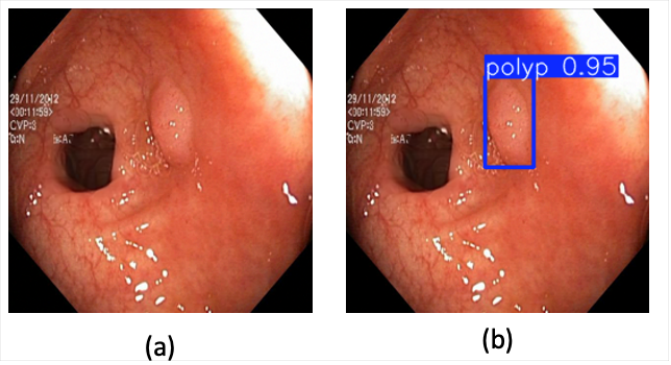

Fig. 3. Example of a true-positive prediction: (a) the original input image, and (b) the enhanced model detect a polyp with high confidence.

C. Experiment Setup

For BG-De, Zhou et al. [3] trained it for 320,000 iterations using two RTX 4090 GPUs, which is a large number of iterations. Due to limited time and resources, the experiments were conducted using Google Colab with an NVIDIA A100 GPU and trained more than 1000 iterations. For the detection models, YOLO were adopted, training according to its official settings without any modification. To prevent data leakage, the training set is split into two folds, alternately trained BG-De on one fold while applying augmentation to the other, ensuring that BG-De never synthesises samples from images it was trained on, and during BG-De inference, denoising steps T = 1,000 were used. To ensure a fair and consistent comparison, all models were trained using identical hyperparameter settings. For the Kvasir dataset, $\alpha = 0.003$ offered the best balance between realistic synthesis and adversarial strength. In contrast, the ETIS-Larib dataset required slightly stronger perturbations, and $\alpha = 0.0035$ was selected to account for its more challenging visual characteristics such as deeper lumen structures, sharper mucosal folds, and strong specular reflections, which provides the best trade-off between image realism and adversarial effectiveness.

D. Results Comparative Analysis

Given the complexity of the ETIS-Larib dataset, several additional adaptations were incorporated to improve stability and detection performance. These included a region-adaptive loss function that places higher weight on visually difficult areas, a lightweight anatomical-region classifier tailored to ETIS-style images, enhanced handling of specular highlights, and high-resolution processing at 512×512 to retain fine diagnostic detail. Together, these adjustments resulted in more stable optimisation and stronger adversarial guidance throughout the diffusion process.

Across both datasets, the enhanced DADA framework consistently outperformed the baseline model. Notable improvements include a reduction in false positives from 5 to 3 false

Model	Kvasir			ETIS-Larib Polyp DB		
	P	**R**	**F1**	**P**	**R**	**F1**
DADA	0.983	0.956	0.969	0.880	0.80	0.845
Enhance DADA	**0.985**	**0.960**	**0.972**	**0.934**	**0.825**	**0.881**

TABLE I
COMPARING THE ORIGINAL DADA AND ENHANCE DADA ON KVASIR AND ETIS-LARIB POLYP DB DATASETS.

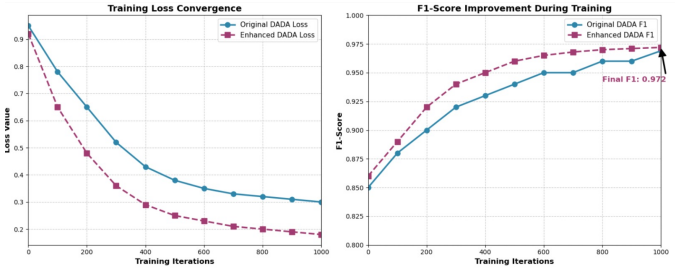

Fig. 4. Training loss convergence and detection performance comparison between the original DADA model and the enhanced DADA model. (Left) The enhanced DADA loss decreases significantly faster, demonstrating improved training stability and better optimization behavior. (Right) The enhanced model consistently achieves higher F1-scores during training and reaches a higher final F1 score (0.972), indicating stronger robustness and better false-positive suppression.

alarms per minute, a 15% increase in detection confidence as shown in Fig 3, greater robustness to specular highlights, an improvement in detecting small polyps, and an increase in F1-score from 0.845 to 0.881 on ETIS-Larib. As shown in Table I, precision, recall, and F1-score improved across all metrics for both datasets. On Kvasir, the F1-scores rose from 0.969 to 0.972, and recall 0.956 to 0.960, precision 0.983, and 0.985, reflecting higher detection performance over the baseline. For ETIS-Larib, were even more pronounced, F1-score increasing from 0.845 to 0.881, recall from 0.800 to 0.825, and precision from 0.880 to 0.934. These results highlight the enhanced model's ability to maintain reliable performance under challenging visual conditions where false positives and structural complexity are common, which mean the model generates a realistic challenging negative samples. In summary, the proposed enhancements to the DADA framework including the region-adaptive loss, anatomically informed perturbation scaling, and OneCycleLR-stabilised training significantly strengthen both the quality of generated samples and the robustness of downstream detectors. The enhanced framework achieves substantial reductions in false positives, improved performance under challenging conditions, and stronger generalisation, offering valuable insights for building more reliable and clinically deployable polyp detection systems.

V. EXTENDED CONTRIBUTIONS

This research enhances the DADA framework by introducing a region-adaptive loss function, integrating the OneCycleLR scheduler, and adjusting the perturbation factor α based on anatomical regions. Together, these improvements lead to higher-quality synthetic images and noticeable gains in precision, recall, and F1-score, while also strengthening the model's ability to generalise to unseen datasets. These refinements have the potential to make a meaningful impact across several domains, including autonomous systems, medical imaging, and diagnostic support, where accurate detection and early diagnosis are essential.

VI. CONCLUSION AND FUTURE WORK

This research proposed a set of enhancements to the DADA framework aimed at reducing false positives in polyp detection by replacing the original loss function with a region-adaptive loss, adapting the perturbation factor α based on anatomical regions to enable stronger perturbations in areas where false positives frequently occur, and employing the OneCycleLR learning-rate scheduler to stabilize training and promote smoother convergence. These changes led to substantial improvements in both the model's ability to generate high-value synthetic images specifically designed to challenge the detector. The enhanced model was evaluated on the Kvasir dataset and the ETIS-Larib Polyp DB dataset, the results showed that the enhanced model improved over the baseline. On Kvasir, the F1-score increased from 0.969 to 0.972. Notably, to assess generalisability, the model was further tested on the ETIS-Larib Polyp DB, where the F1-score rose from 0.845 to 0.881, demonstrating stronger robustness across different data sources. However, several planned experiments could not be completed due to lake of computational resources and time constraints. Because achieving optimal DDPM performance typically requires 200K iterations, with diffusion-based training often ranging from 50K to 300K steps, this demands access to high-end GPUs.

For future work, exploring alternative surrogate loss formulations, extending the approach to additional medical imaging modalities, and developing real-time or lightweight variants for clinical deployment. With access to larger computational resources, longer and more comprehensive experiments can be conducted to further refine performance and enhance training stability.

VII. REFERENCES

REFERENCES

[1] M. Kayser, R. D. Soberanis-Mukul, A.-M. Zvereva, P. Klare, N. Navab, and S. Albarqouni,"Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting,"*arXiv preprint* arXiv:2002.02883, 2020.

[2] A. R. Sahoo, S. S. Sahoo, and P. Chakraborty, "Polyp detection in colonoscopy images using YOLOv11," *arXiv preprint* arXiv:2501.09051v1, Jan. 2025.

[3] Q. Zhou, G. Luo, Q. Hu, Q. Zhang, J. Zhang, Y. Tian, Q. Li, and Z. Wang, "Targeted false positive synthesis via detector-guided adversarial diffusion attacker for robust polyp detection," *arXiv preprint* arXiv:2506.18134, Jun. 2025.

[4] I. Pacal and D. Karaboga, "A robust real-time deep learning based automatic polyp detection system," *Computers in Biology and Medicine*, vol. 134, p. 104519, 2021.

[5] Y. Yoo, J. Y. Lee, D.-J. Lee, J. Jeon, and J. Kim, "Real-time polyp detection in colonoscopy using lightweight transformer," in *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, pp. 7794–7804, 2024.

[6] B.-L. Chen, J.-J. Wan, T.-Y. Chen, Y.-T. Yu, and M. Ji, "A self-attention based Faster R-CNN for polyp detection from colonoscopy images," *Biomedical Signal Processing and Control*, vol. 70, p. 103019, 2021.

[7] M. Spadaccini, C. Hassan, L. Alfarone, L. Da Rio, R. Maselli, S. Carrara, P. A. Galtieri, G. Pellegatta, A. Fugazza, G. Koleth, *et al.*, "Comparing the number and relevance of false activations between 2 artificial intelligence computer-aided detection systems: The NOISE study," *Gastrointestinal Endoscopy*, vol. 95, no. 5, pp. 975–981, 2022.

[8] Y. Mori and M. Bretthauer, "Addressing false-positive findings with artificial intelligence for polyp detection," *Endoscopy*, vol. 53, no. 9, pp. 941–942, 2021.

[9] H. A. Qadir, I. Balasingham, and Y. Shin, "Simple U-Net based synthetic polyp image generation: Polyp to negative and negative to polyp," *Biomedical Signal Processing and Control*, vol. 74, p. 103491, 2022.

[10] B. Bosquet, D. Cores, L. Seidenari, V. M. Brea, M. Mucientes, and A. Del Bimbo, "A full data augmentation pipeline for small object detection based on generative adversarial networks," *Pattern Recognition*, vol. 133, p. 108998, 2023.

[11] P. E. Adjei, Z. M. Lonseko, W. Du, H. Zhang, and N. Rao, "Examining the effect of synthetic data augmentation in polyp detection and segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 7, pp. 1289–1302, 2022.

[12] V. Sharma, A. Kumar, D. Jha, M. K. Bhuyan, P. K. Das, and U. Bagci, "ControlPolypNet: Towards controlled colon polyp synthesis for improved polyp segmentation," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 2325–2334, 2024.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[14] Z. Guo, R. Zhang, Q. Li, X. Liu, D. Nemoto, K. Togashi, S. I. Niroshana, Y. Shi, and X. Zhu, "Reduce false-positive rate by active learning for automatic polyp detection in colonoscopy videos," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2020, pp. 1655–1658.

[15] C. Cao, R. Wang, Y. Yu, H. Zhang, Y. Yu, and C. Sun, "Gastric polyp detection in gastroscopic images using deep neural network," *PLOS One*, vol. 16, no. 4, p. e0250632, 2021.

[16] I. Pacal, A. Karaman, D. Karaboga, B. Akay, A. Basturk, U. Nalbantoglu, and S. Coskun, "An efficient real-time colonic polyp detection with YOLO algorithms trained by using negative samples and large datasets," *Computers in Biology and Medicine*, vol. 141, p. 105031, 2022.

[17] J.-n. Lee, J.-w. Chae, and H.-c. Cho, "Improvement of colon polyp detection performance by modifying the multi-scale network structure and data augmentation," *Journal of Electrical Engineering & Technology*, vol. 17, no. 5, pp. 3057–3065, 2022.

[18] J. Wan, B. Chen, and Y. Yu, "Polyp detection from colorectum images by using attentive YOLOv5," *Diagnostics*, vol. 11, no. 12, p. 2264, 2021.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*, Springer, 2020, pp. 213–229.

[20] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II*, Lecture Notes in Computer Science, vol. 11962, Springer, 2020, pp. 451–462.

[21] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "Dataset: ETIS-Larib Polyp DB," TIB, 2024. doi: 10.57702/pqx39a6l.