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

Quan Zhou[†1], Gan Luo[†1], Qiang Hu[(✉)2], Qingyong Zhang[1], Jinhua Zhang[3], Yinjiao Tian[3], Qiang Li[2], and Zhiwei Wang[(✉)2]

[1] Wuhan University of Technology
[2] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology
[3] Changzhou United lmaging Healthcare Surgical Technology Co., Ltd.
{zhouquan910, ganluo}@whut.edu.cn, {huqiang77, zwwang}@hust.edu.cn

Abstract. Polyp detection is crucial for colorectal cancer screening, yet existing models are limited by the scale and diversity of available data. While generative models show promise for data augmentation, current methods mainly focus on enhancing polyp diversity, often overlooking the critical issue of false positives. In this paper, we address this gap by proposing an adversarial diffusion framework to synthesize high-value false positives. The extensive variability of negative backgrounds presents a significant challenge in false positive synthesis. To overcome this, we introduce two key innovations: First, we design a regional noise matching strategy to construct a negative synthesis space using polyp detection datasets. This strategy trains a negative-centric diffusion model by masking polyp regions, ensuring the model focuses exclusively on learning diverse background patterns. Second, we introduce the Detector-guided Adversarial Diffusion Attacker (DADA) module, which perturbs the negative synthesis process to disrupt a pre-trained detector's decision, guiding the negative-centric diffusion model to generate high-value, detector-confusing false positives instead of low-value, ordinary backgrounds. Our approach is the first to apply adversarial diffusion to lesion detection, establishing a new paradigm for targeted false positive synthesis and paving the way for more reliable clinical applications in colorectal cancer screening. Extensive results on public and in-house datasets verify the superiority of our method over the current state-of-the-arts, with our synthesized data improving the detectors by at least 2.6% and 2.7% in F1-score, respectively, over the baselines. Codes are at https://github.com/Huster-Hq/DADA.

Keywords: Adversarial diffusion framework · Data synthesis · Colorectal polyp detection.

---

† Equal contribution; ✉ corresponding author.

1   Introduction

Colorectal cancer (CRC) remains one of the most prevalent and deadly malignancies worldwide, presenting significant challenges to global public health [1]. Early and accurate screening of colonic polyps during colonoscopy is critical in reducing the morbidity and mortality associated with CRC. However, polyp screening is often hindered by the variability in clinicians' skills and experience levels, which underscores the urgent need for automatic polyp detection.

Recent advancements in polyp detection have primarily focused on innovations in network architectures, such as Transformer-based [2], YOLO-based [3], R-CNN-based [4,5]. Further progress has been made by leveraging temporal relationships to improve detection accuracy, utilizing techniques like 3D models [6], optical flow correlations [7], and tracking modules [8]. Despite these advances, polyp detection still faces substantial challenges due to the variability in polyp morphology and the complex, dynamic nature of the colon environment. While certain methods address specific issues, such as reflection artifacts [9,10] and size variability [11], the lack of comprehensive and diverse datasets remains a major barrier to further progress [12].

In recent years, breakthroughs in generative models, particularly denoising diffusion probabilistic models (DDPMs), have provided new opportunities to enhance object detection by generating challenging, high-value training data. For example, DS-GAN [13] has proven effective in generating realistic small targets, addressing the challenges of detecting small objects. In polyp detection, Qadir *et al.* [14] introduced a conditional GAN-based framework for synthesizing synthetic polyp images, improving both detection and segmentation accuracy. Similarly, Adjei *et al.* [12] integrated a modified pix2pix framework with traditional augmentation techniques to boost model performance. More recently, ControlPolypNet [15] employed user-controllable inputs to generate clinically relevant polyp images, expanding datasets and improving segmentation outcomes.

However, these methods primarily focus on generating positive samples, with relatively little attention to the issue of **false positives**. In a 15- to 20-minute polyp screening, negative samples significantly outnumber positive ones, with current systems producing an average of five false positives per minute [16]. This excessive number of false positives can disrupt clinicians' focus [17], a phenomenon known as the 'crying wolf effect'. It not only increases the workload of clinicians but also introduces the risk of unnecessary medical interventions and costly diagnostic errors [18]. Furthermore, individual variability and differences in bowel preparation result in highly complex colon environments, further exacerbating the false positive problem and making reliable detection more difficult.

In this paper, we address a critical gap for the first time by proposing a novel method for targeted false positive generation. Our approach seeks to reduce false positives by generating challenging negative training samples that mimic polyp-like features, which can confuse the detector. Specifically, our method consists of two key components: a background-only DDPM as the base negative generator and a Detector-guided Adversarial Diffusion Attacker (DADA) module. The base negative generator is trained on *off-the-shelf* polyp detection datasets but

employs a regional noise matching strategy to decouple the polyp regions from the DDPM's training. This strategy ensures that the generator focuses solely on learning diverse non-polyp background visual patterns to avoid leak of polyp information into the subsequent generation of polyp-like interferences. During the denoising process, the DADA module, inspired by adversarial attack principles, establishes a gradient backpropagation pathway to enable the injection of adversarial perturbations into the denoising process, effectively confusing the detector's decision boundary and guiding the generation of general low-value negatives into high-value false positive samples.

In summary, our major contributions are as follows:

– We propose a novel image synthesis method that integrates diffusion models with adversarial attacks, focusing on generating high-value negative samples capable of effectively misleading polyp detectors.
– We propose a background-only denoiser learning to generate pure negative patterns from *off-the-shelf* polyp detection datasets, and a Detector-guided Adversarial Diffusion Attacker (DADA) module guiding the denoising to high-value realistic false positive ones by attacking a well-trained detector.
– Extensive experiments demonstrate that our method achieves state-of-the-art performance on both the public Kvasir and our in-house dataset, with improvements of at least 2.6% and 2.7% in F1-score, respectively, over the baselines.

2 Method

Fig. 1 shows the overall of our adversarial diffusion framework, which mainly consists of three key modules: (1) a well-trained polyp detector, (2) a background-only denoiser (BG-De), and (3) a detector-guided adversarial diffusion attacker (DADA). During inference, the iterative process alternates between denoising through BG-De, evaluating their potential for false positive induction via the detector's prediction, and applying DADA-computed adversarial perturbations to confuse the detector in a pre-defined region of interest.

2.1 Background-only Denoiser as Base Negative Generator using Regional Noise Matching

BG-De is a variant of the Diffusion Probabilistic Model (DDPM) [19], designed to selectively denoise background regions. During the forward diffusion process, BG-De adds Gaussian noise to an image x_0 over t time steps: $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon_{x_t}$, where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\alpha_{1:T}$ are hyperparameters controlling the signal-to-noise ratio, and $\epsilon_{x_t} \sim \mathcal{N}(0, 1)$. BG-De is trained to reverse this process by predicting the noise $\epsilon_\theta(x_t, t)$, enabling the reconstruction of the image. Unlike the original DDPM, BG-De uses regional masking to focus learning on background regions. For a polyp detection training sample $\{x, m_{gtb}\}$, where m_{gtb} is a binarized mask indicating the ground truth (GT) bounding boxes of polyps
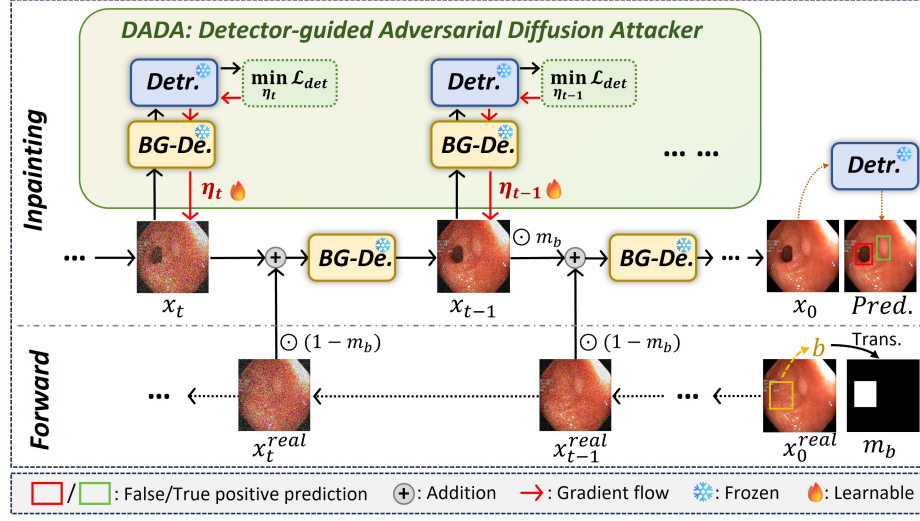
Fig. 1. The inference pipeline consists of three key modules: the background-only denoiser (BG-De), a well-trained polyp detector (Detr), and the detector-guided adversarial diffusion attacker (DADA). Inpainting is used to generate new negative samples in a user-specified region of the real image. BG-De samples from the pure *negative* distribution, while DADA guides the sampling toward visual patterns that elicit a *positive* response from the detector. Note that the training of both BG-De and the detector is not shown in this figure.

(with 1 assigned to pixels inside the bounding boxes and 0 outside), the loss function is modified as:

$$\mathcal{L}_{\mathrm{BG}} = \mathbb{E}_{x_0,\epsilon,t}\|(1 - m_{gtb}) \odot (\epsilon_{x_t} - \epsilon_\theta(x_t,t))\|^2. \tag{1}$$

This approach allows BG-De to only model the background (negative) sample distribution, even when polyp samples are present. Notably, the training sets consist solely of polyp images and their corresponding GT bounding boxes, which are standard in detection tasks and *do not require extra data collection efforts*.

2.2 Detector-guided Adversarial Diffusion Attacker Guiding Ordinary Backgrounds to High-value False Positives

Building on the trained BG-De, we progressively denoise an initialized noise map to synthesize realistic negative endoscopic images x_0. The denoising process is formulated as:

$$x_{t-1} = \mu_\theta(x_t,t) + \sqrt{\beta_t}\epsilon, \tag{2}$$

where $\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right)$ is the sampled mean value map by BG-De, and $\epsilon \sim \mathcal{N}(0,1)$ is for reparameterization.

However, the generated images are typically ordinary and insufficient to confuse the detector, limiting their training value. To address this, we introduce a perturbation η_t at each denoising step to modify the sampling trajectory and guide BG-De toward generating images that exceed the detector's decision boundary. The denoising process becomes:

$$x_{t-1} = \mu_\theta(x_t + \eta_t, t) + \sqrt{\beta_t}\epsilon, \tag{3}$$

where η_t is a trainable variable initialized as a *zero* matrix.

Inspired by adversarial attack methods [20,21], we propose the DADA module to optimize the perturbation by targeting the well-trained detector, aiming to induce false positives in a user-defined bounding box b. This generates challenging negative samples with misleading interference features. Specifically, we feed x_{t-1} into the detector, treating b as an "illusory" ground truth (GT) box, and calculate the detection loss as:

$$\mathcal{L}_{det} = \mathcal{L}_{cls.}(1, \hat{p}_{\sigma(b)}) + \mathcal{L}_{loc.}(b, \hat{b}_{\sigma(b)}), \tag{4}$$

where $\mathcal{L}_{cls.}$ and $\mathcal{L}_{loc.}$ are classification and localization losses, respectively. $\hat{p}_{\sigma(b)}$ and $\hat{b}_{\sigma(b)}$ represent the predicted polyp-class probability and bounding box for the "illusory" GT box b. The prediction-GT assignment is related to the detector, *e.g.* Hungarian algorithm [22] in DETR [23]. The perturbation is then optimized by minimizing the detection loss, and the direction of η_t is updated via backpropagated gradients:

$$\eta_t = \eta_t - \alpha \cdot \text{sgn}(\bigtriangledown_{\eta_t} \mathcal{L}_{det}), \tag{5}$$

where $\text{sgn}(\cdot)$ is the signum function and α is a small step size. By integrating DADA with BG-De, perturbations at each denoising step guide the generation of x_0, creating high-value negative samples that significantly confuse the detector.

2.3   Inpainting Strategy to Maintain Context Consistency

To enhance the fidelity of the generated image regarding anatomical structures, we use an inpainting strategy that incorporates the real image context to guide local false positive generation. Specifically, we apply "attack & inpaint" within the pre-defined region, while the rest of the regions are kept as the noisy version of the real image. We define the inpainting region as the same as b described in the Sec. 2.2. The final denoising process is formulated as follows:

$$x_{t-1} = \mu_\theta\left((1 - m_b) \odot x_t^{real} + m_b \odot (x_t + \eta_t), t\right) + \sqrt{\beta_t}\epsilon, \tag{6}$$

where m_b is a binarized map, with 1 assigned to pixels within b and 0 outside. This strategy integrates DADA with BG-De, introducing perturbations at each denoising step to generate x_0, which maximizes detector confusion while preserving visual integrity. At last, our method can generate high-value negative samples in a localized region of the real image while maintaining overall context, as shown in Fig. 2.

Table 1. Quantitative comparison on Kvasir and our private dataset. The best performance is marked in bold.

Backbone	Methods	Kvasir			In-house		
		P	R	F1	P	R	F1
YOLO [25]	baseline	0.941	0.940	0.941	0.894	0.813	0.852
	APGD [26]	0.952	0.942	0.947	0.905	0.815	0.858
	FAB [27]	0.947	0.944	0.945	0.901	0.817	0.857
	Repaint [28]	0.951	0.946	0.948	0.911	0.823	0.865
	LaMa [29]	0.953	0.947	0.950	0.915	0.819	0.864
	Ours	**0.983**	**0.956**	**0.969**	**0.942**	**0.845**	**0.891**
DETR [23]	baseline	0.950	0.856	0.901	0.873	0.574	0.693
	APGD [26]	0.957	0.861	0.906	0.879	0.576	0.696
	FAB [27]	0.955	0.858	0.904	0.876	0.577	0.696
	Repaint [28]	0.962	0.862	0.909	0.885	0.582	0.702
	LaMa [29]	0.963	0.859	0.908	0.886	0.585	0.705
	Ours	**0.980**	**0.880**	**0.927**	**0.918**	**0.592**	**0.720**

3 Experiments
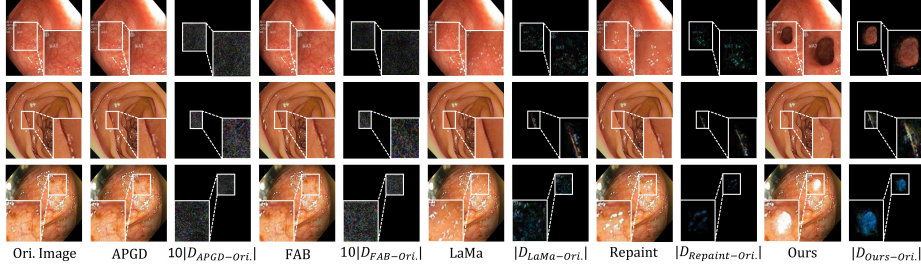
3.1 Datasets and Evaluation Metrics

We conduct experiments on two datasets: Kvasir [24] and an in-house dataset. The Kvasir dataset consists of $1,000$ polyp images, where we generate GT boxes from the provided GT masks. The in-house dataset includes $1,516$ polyp images with GT boxes carefully labeled by two experienced endoscopists from a local hospital. For both datasets, we randomly split them into training, validation, and test sets with an $8:1:1$ ratio. We evaluate detection performance using Precision (P), Recall (R), and F1-score (F1).

3.2 Implementation Details

For BG-De, we implement the model architecture based on the conventional DDPM [19] and train it for $320,000$ iterations on two RTX 4090 GPUs with a batch size of 20. We use YOLO [25] and DETR [23] as detection models, training them according to their official settings. We split the training set into two folds, alternately training BG-De on one while augmenting the other. For BG-De inference, we set the denoising steps to $1,000$, the final image size to 256×256, and the perturbation step size α in Eq. (5) to 0.003.

3.3 Comparison with State-of-the-arts

We compare our method with four state-of-the-art (SOTA) methods on both Kvasir and the in-house dataset: two adversarial attack methods (APGD [26], FAB [27]) and two inpainting methods (Repaint [28], LaMa [29]). To ensure fairness, we implement all methods as described in Sec. 3.2, augment each image once, and train both YOLO and DETR with the original and augmented datasets. Baselines are trained only on the original training set.

Ori. Image | APGD | $10|D_{APGD-Ori.}|$ | FAB | $10|D_{FAB-Ori.}|$ | LaMa | $|D_{LaMa-Ori.}|$ | Repaint | $|D_{Repaint-Ori.}|$ | Ours | $|D_{Ours-Ori.}|$

Fig. 2. Visualizations of synthesis images by different methods. The white bounding boxes indicate pre-defined attack/inpainting regions. For clarity, we visualize difference maps D between synthesis images and original images.

Table 2. Ablation study of two key components, *i.e.*, BG-De. and DADA.

Backbone	Components		Kvasir			In-house		
	DADA	BG-De	P	R	F1	P	R	F1
YOLO	✗	✗	0.936	0.939	0.937	0.891	0.812	0.850
	✓	✗	0.933	0.930	0.931	0.882	0.801	0.840
	✗	✓	0.950	0.945	0.947	0.902	0.818	0.858
	✓	✓	**0.983**	**0.956**	**0.969**	**0.942**	**0.845**	**0.891**
DETR	✗	✗	0.945	0.853	0.897	0.870	0.571	0.689
	✓	✗	0.945	0.843	0.891	0.865	0.565	0.684
	✗	✓	0.952	0.861	0.904	0.877	0.581	0.699
	✓	✓	**0.980**	**0.880**	**0.927**	**0.918**	**0.592**	**0.720**

As shown in Table 1, our method significantly improves the performance of both YOLO and DETR across the two datasets, outperforming all baselines. This improvement mainly stems from our methods ability to generate high-value negative samples. As shown in Fig. 2, our method generates challenging negative samples such as circular lumens, colonic folds, and specular highlights, which are likely to mislead the baseline model into producing false positives. Using these samples as additional training data greatly enhances the performance, especially in terms of precision. In contrast, methods like APGD and FAB generate subtle, noise-like perturbations without clear semantics, while Repaint and LaMa produce ordinary backgrounds that offer limited benefits.

3.4 Ablation Study

To verify the effectiveness of the two key components, *i.e.*, BG-De and DADA, we train three variants of YOLO and DETR by disabling BG-De and/or DADA. The comparison results of these methods on Kvasir and our in-house dataset are listed in Table 2. From the first two rows of each detector, we observe that disabling BG-De results in performance degradation compared to their baselines in Table 1. This is primarily because the generator randomly generates both

Table 3. Ablation on perturbation step size α. $\uparrow$ and $\downarrow$ denote 'higher is better' and 'lower is better', respectively.

α	Kvasir			In-house		
	FID $\downarrow$	FPGR $\uparrow$	F1 $\uparrow$	FID $\downarrow$	FPGR $\uparrow$	F1 $\uparrow$
0.001	**4.62**	0.177	0.959	**2.78**	0.216	0.868
0.002	5.97	0.681	0.967	3.12	0.645	0.887
0.003	8.23	0.885	**0.969**	3.53	0.849	**0.891**
0.004	9.64	0.924	0.963	3.93	0.934	0.861
0.005	10.75	**0.951**	0.954	4.20	**0.965**	0.859

background and polyps, and treating all generated samples as backgrounds introduces label noise and compromises detector training. As shown in the third row of each detector, introducing BG-De alone provides limited performance improvement, which is mainly because most of the negative samples generated by BG-De are of low-value to the detectors. Finally, when DADA and BG-De are jointly utilized, both YOLO and DETR achieve the best performance with significant improvements, which demonstrates that DADA and BG-De are highly complementary and must be used in conjunction.

3.5   Hyperparameter Choices

To explore the impact of different step sizes α in Eq. (5) on detector performance, we conduct a hyperparameter search using YOLO on both the Kvasir and in-house datasets. We introduce two additional metrics: Fréchet Inception Distance (FID) and False Positive Generation Rate (FPGR). FID measures the distribution discrepancy between generated and real images, while FPGR quantifies the proportion of generated images causing false positives. As shown in Table 3, increasing α results in higher FPGR but lower FID, as larger α values introduce more noise into generated images. Based on these results, we set $\alpha = 0.003$, which yields the best detector performance with an F1-score of 0.969 on Kvasir and 0.891 on the in-house dataset.

4   Conclusion

In this work, we introduce a novel approach for generating high-value false positive samples to address the challenge of limited data diversity in polyp detection systems. By integrating a feedback mechanism that targets the detector within the denoising process of a diffusion model, we are able to synthesize realistic, challenging negative samples. This approach, which is the first of its kind in the context of polyp detection, allows for a significant improvement in detector performance by augmenting the training set with these strategically generated false positives. Extensive experiments on both the public Kvasir dataset and our in-house dataset demonstrate that our method enables detectors like YOLO and DETR to achieve state-of-the-art results. Our work contributes a new paradigm

for data synthesis in clinical applications, particularly in colorectal cancer screening, where diverse training data is critical for improving detection accuracy and reducing false positives.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Juan José Granados-Romero, Alan Isaac Valderrama-Treviño, Ericka Hazzel Contreras-Flores, Baltazar Barrera-Mera, Miguel Herrera Enríquez, Karen Uriarte-Ruíz, Jesús Carlos Ceballos-Villalba, Aranza Guadalupe Estrada-Mata, Cristopher Alvarado Rodríguez, and Gerardo Arauz-Peña. Colorectal cancer: a review. *Int J Res Med Sci*, 5(11):4667, 2017.
2. Youngbeom Yoo, Jae Young Lee, Dong-Jae Lee, Jiwoon Jeon, and Junmo Kim. Real-time polyp detection in colonoscopy using lightweight transformer. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7794–7804. IEEE, 2024.
3. Ishak Pacal and Dervis Karaboga. A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*, 134:104519, 2021.
4. Xiaoyong Yang, Qianxing Wei, Changhe Zhang, Kaibo Zhou, Li Kong, and Weiwei Jiang. Colon polyp detection and segmentation based on improved mrcnn. *IEEE Transactions on Instrumentation and Measurement*, 70:1–10, 2020.
5. Bo-Lun Chen, Jing-Jing Wan, Tai-Yue Chen, Yong-Tao Yu, and Min Ji. A self-attention based faster r-cnn for polyp detection from colonoscopy images. *Biomedical Signal Processing and Control*, 70:103019, 2021.
6. Juana González-Bueno Puyal, Patrick Brandao, Omer F Ahmad, Kanwal K Bhatia, Daniel Toth, Rawen Kader, Laurence Lovat, Peter Mountney, and Danail Stoyanov. Polyp detection on video colonoscopy using a hybrid 2d/3d cnn. *Medical Image Analysis*, 82:102625, 2022.
7. He Zheng, Hanbo Chen, Junzhou Huang, Xuzhi Li, Xiao Han, and Jianhua Yao. Polyp tracking in video colonoscopy using optical flow with an on-the-fly trained cnn. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 79–82. IEEE, 2019.
8. Tao Yu, Ne Lin, Xu Zhang, Yanqi Pan, Huiyi Hu, Wenfang Zheng, Jiquan Liu, Weiling Hu, Huilong Duan, and Jianmin Si. An end-to-end tracking method for polyp detectors in colonoscopy videos. *Artificial intelligence in medicine*, 131:102363, 2022.
9. Maxime Kayser, Roger D Soberanis-Mukul, Anna-Maria Zvereva, Peter Klare, Nassir Navab, and Shadi Albarqouni. Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting. *arXiv preprint arXiv:2002.02883*, 2020.

10. Qiang Hu, Zhenyu Yi, Ying Zhou, Fang Peng, Mei Liu, Qiang Li, and Zhiwei Wang. Sali: Short-term alignment and long-term interaction network for colonoscopy video polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 531–541. Springer, 2024.

11. Meryem Souaidi and Mohamed El Ansari. Multi-scale hybrid network for polyp detection in wireless capsule endoscopy and colonoscopy images. *Diagnostics*, 12(8):2030, 2022.

12. Prince Ebenezer Adjei, Zenebe Markos Lonseko, Wenju Du, Han Zhang, and Nini Rao. Examining the effect of synthetic data augmentation in polyp detection and segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 17(7):1289–1302, 2022.

13. Brais Bosquet, Daniel Cores, Lorenzo Seidenari, Víctor M Brea, Manuel Mucientes, and Alberto Del Bimbo. A full data augmentation pipeline for small object detection based on generative adversarial networks. *Pattern Recognition*, 133:108998, 2023.

14. Hemin Ali Qadir, Ilangko Balasingham, and Younghak Shin. Simple u-net based synthetic polyp image generation: Polyp to negative and negative to polyp. *Biomedical Signal Processing and Control*, 74:103491, 2022.

15. Vanshali Sharma, Abhishek Kumar, Debesh Jha, MK Bhuyan, Pradip K Das, and Ulas Bagci. Controlpolypnet: Towards controlled colon polyp synthesis for improved polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2325–2334, 2024.

16. Marco Spadaccini, Cesare Hassan, Ludovico Alfarone, Leonardo Da Rio, Roberta Maselli, Silvia Carrara, Piera Alessia Galtieri, Gaia Pellegatta, Alessandro Fugazza, Glenn Koleth, et al. Comparing the number and relevance of false activations between 2 artificial intelligence computer-aided detection systems: the noise study. *Gastrointestinal endoscopy*, 95(5):975–981, 2022.

17. Erik A Holzwanger, Mohammad Bilal, Jeremy R Glissen Brown, Shailendra Singh, Aymeric Becq, Kenneth Ernest-Suarez, and Tyler M Berzin. Benchmarking definitions of false-positive alerts during computer-aided polyp detection in colonoscopy. *Endoscopy*, 53(09):937–940, 2021.

18. Yuichi Mori and Michael Bretthauer. Addressing false-positive findings with artificial intelligence for polyp detection. *Endoscopy*, 53(09):941–942, 2021.

19. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

20. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

21. Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

22. Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

23. Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

24. 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 26*, pages 451–462. Springer, 2020.

25. Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Piotr Skalski, Adam Hogan, et al. ultra-lytics/yolov5: v6. 0-yolov5n'nano'models, roboflow integration, tensorflow export, opencv dnn support. *Zenodo*, 2021.
26. Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.
27. Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, pages 2196–2205. PMLR, 2020.
28. Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.
29. Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.