# 修 士 論 文 概 要 書

Master's Thesis Summary

Date of submission: 01/27/2025 (MM/DD/YYYY)

| | | | | | |
|---|---|---|---|---|---|
| 専攻名 (専門分野)<br>Department | Computer Science and Communications Engineering | 氏 名<br>Name | Thao Phuong Nguyen | 指 導 教 員<br>Advisor | Hiroshi Watanabe<br>印<br>Seal |
| 研究指導名<br>Research guidance | Audiovisual Information Processing | 学籍番号<br>Student ID number | CD<br>5123FG05-4 | | |
| 研究題目<br>Title | Real-time Polyp Detection using Deep Learning on Endoscopic Images and Videos | | | | |

## 1. Introduction

Endoscopy plays a vital role in clinical practice by enabling the examination of internal cavities within the human body. In medicine, endoscopy is classified based on the organ being investigated. It serves as the primary method for identifying abnormalities, including both precancerous and cancerous conditions in the gastrointestinal tract, with diagnoses confirmed through biopsy sampling. Various endoscopic techniques are available for gastrointestinal tract examinations, offering detailed insights into abnormal areas.

Despite considerable progress, achieving accurate and real-time automatic polyp detection remains a significant challenge. This is primarily due to the wide variations in polyp characteristics, such as shape, texture, size, and color, as well as the presence of artifacts that are similar to polyps during endoscopy.

In this study, we propose a novel Gaussian Enhanced Euclidean norm Ghost attention (GEEG) module to enable reliable real-time polyp detection in endoscopic images and videos. This innovative attention mechanism enhances the features generated by the Ghost convolution's [1] cheap operations by improving the ability in the extraction of inter-channel and spatial information within the convolutional layers. The proposed module is incorporated into the YOLOv8 [2] backbone, creating a new model named GEEG-YOLOv8, aimed at overcoming the obstacles in polyp detection. Experimental evaluations on three public datasets reveal that our method surpasses current state-of-the-art approaches in both accuracy and detection speed.

## 2. Proposed Method

The overall framework of GEEG-YOLOv8 is shown in Fig. 1. The architecture is composed of a backbone, a neck, and a detection head. The backbone integrates several regular GhostConv [1] and GEEG-C2f modules. While the GEEG module improves information extraction capability, it also adds complexity to the model. To address this, GEEG is only incorporated into C2f to minimize the addition of excessive parameters.

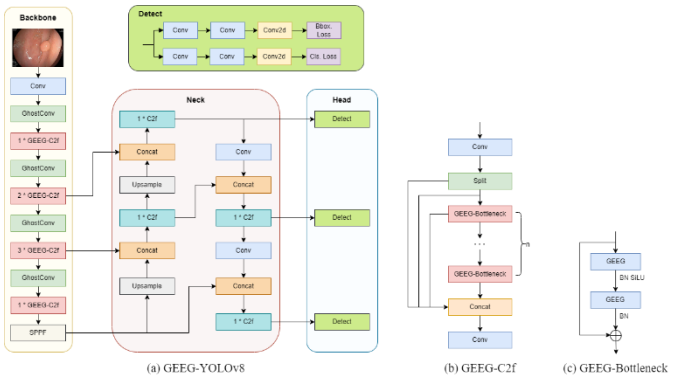



Fig. 1. The overall architecture of the proposed GEEG-YOLOv8. (Reproduced from our previous work [3].)

The GEEG-Bottleneck, illustrated in Fig. 1c, consists of two GEEG modules. The first GEEG module acts as a squeezing layer, reducing the input channels by half, while the second GEEG module expands the channels to match the shortcut path. Then the inputs and outputs of these two GEEG modules are connected by a residual connection. The first GEEG module is followed by Batch Normalization (BN) [4] and the SiLU [5] activation function, while only BN is applied after the second module. The GEEG-C2f, depicted in Fig. 1b, includes $n$ GEEG-Bottleneck with two parallel gradient flow branches, allowing it to capture richer gradient flow information while reducing parameters. The neck utilizes PANet, which combines feature maps from lower layers with deeper layers to enhance information flow.

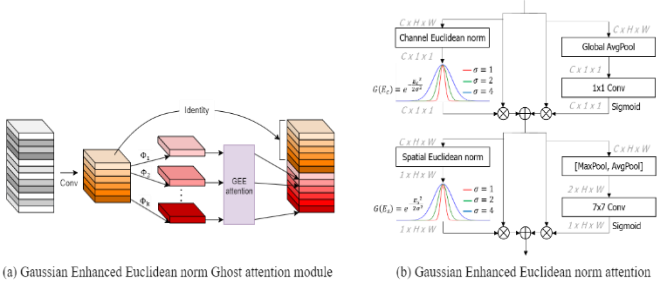



Fig. 2. Gaussian Enhanced Euclidean norm Ghost attention module block details. (Reproduced from our previous work [3].)

GEEG-YOLOv8 adopts an anchor-free approach with a decoupled detection head. The first branch of the detection head outputs the loss for object bounding box, while the second branch outputs the loss for object classification.

Self-attention [6] has been proven to be highly effective in capturing long-range global dependencies. However, its quadratic computational complexity makes it impractical for real-time deployment of hardware with limited resources. To address this challenge, we introduce the lightweight Gaussian Enhanced Euclidean norm (GEE) attention mechanism, which is designed to enhance GhostConv performance with a negligible number of additional parameters. This approach is based on the hypothesis that smaller attention activations are linked to global contexts with larger absolute values [7]. GEE consists of two modules: channel attention and spatial attention [8], each featuring two branches, as depicted in Fig. 2b.

GEE attention mechanism is integrated after cheap operations in GhostConv to form GEEG module, as depicted in Fig. 2a. Although conventional GhostConv can reduce computational cost, its capacity for information extraction is limited, as the cheap operations only capture local information from inherent feature maps produced by $1 \times 1$ point-wise convolution. By incorporating GEE attention mechanism, GhostConv's ability to capture long-range global channel and spatial information is significantly elevated.

## 3. Experiments

All experiments were performed using an NVIDIA RTX 4070 GPU with 12GB of VRAM. We evaluate our model on various polyp detection datasets. Overall, our model outperforms other methods across all metrics on many datasets. The proposed GEEG-YOLOv8 also outperforms other models specifically designed for polyp detection. Thanks to GEEG's efficient and lightweight design, the number of parameters and GFLOPs in our model are reduced compared to the original YOLOv8. While it does not achieve the lowest number of parameters and GFLOPs, GEEG-YOLOv8 obtains the second-highest FPS at 303. Our proposed method achieves new state-of-the-art in polyp detection accuracy with low model complexity, showing its potential for real-world applications where real-time processing is essential. Qualitative results are shown in Fig. 3.

## 4. Conclusion

This thesis proposes a novel Gaussian Enhanced Euclidean norm Ghost attention module designed to enhance the accuracy and speed of polyp detection in endoscopic images and videos.

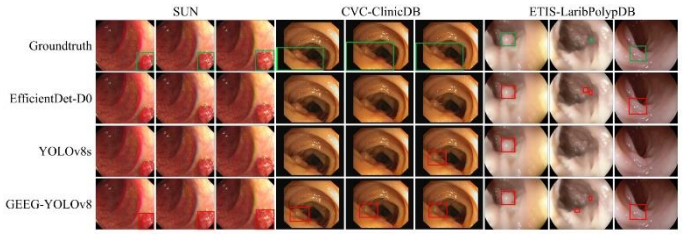



Fig. 3. Qualitative comparisons of polyp detection on three datasets. Green and red denote the ground truth and model prediction, respectively. (Reproduced from our previous work [3].)

The module utilizes a Gaussian function applied to the Euclidean norm of channel and spatial dimension, refining the output features of channel and spatial attention mechanism. This approach improves Ghost convolution's ability to capture long-range global contextual information. The module is incorporated into the backbone of YOLOv8, forming a new model called GEEG-YOLOv8. Extensive experimental evaluations show that the proposed method achieves robust generalization without requiring information from additional dimensions. With a minimal increase in model parameters and GFLOPs, GEEG-YOLOv8 achieves state-of-the-art performance in polyp detection across three datasets.

## References

[1] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[2] J. Glenn, *Ultralytics YOLO,* 2023.

[3] P. T. Nguyen and H. Watanabe, "GEEG-YOLOv8: Gaussian Enhanced Euclidean Norm Ghost Attention for Real-Time Polyp Detection," in *2024 IEEE International Conference on Image Processing (ICIP)*, 2024.

[4] S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167,* 2015.

[5] S. Elfwing, E. Uchibe and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural networks,* vol. 107, pp. 3-11, 2018.

[6] D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929,* 2020.

[7] D. Ruan, D. Wang, Y. Zheng, N. Zheng and M. Zheng, "Gaussian context transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[8] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018.

# Real-time Polyp Detection using Deep Learning on Endoscopic Images and Videos

A Thesis Submitted to the Department of Computer Science and Communications Engineering, the Graduate School of Fundamental Science and Engineering of Waseda University in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: January 27th, 2025

Thao Phuong Nguyen

(5123FG05-4)

Advisor: Prof. Hiroshi Watanabe

Research guidance: Research on Audiovisual Information Processing

## ACKNOWLEDGEMENTS

I respectfully take this opportunity to acknowledge many people who deserve a special mention for their varied contributions during my graduation research. I could never have achieved this work without their kind support and encouragement.

I would like to express my deepest sense of gratitude to my supervisor Professor Hiroshi Watanabe for being an outstanding advisor and excellent mentor. His constant encouragement, patient guidance, support in various ways and invaluable suggestions made this work successful. It was a great privilege and honor to work and study under his guidance.

I would truly like to thank my mom for her overwhelming affection and devotion that assisted me through every aspect of my life.

I wish to thank my friends for always encouraging me to finish this work and for understanding how important it is to me.

I dedicate this work to each and every member of my dearest in an attempt to express my gratitude for their priceless sacrifices.

Finally, I thank myself for putting incredible effort and time into completing the graduation thesis successfully.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## ABSTRACT

Endoscopy plays a vital role in clinical practice by enabling the examination of internal cavities within the human body. In medicine, endoscopy is classified based on the organ being investigated. It serves as the primary method for identifying abnormalities, including both precancerous and cancerous conditions in the gastrointestinal tract, with diagnoses confirmed through biopsy sampling. Various endoscopic techniques are available for gastrointestinal tract examinations, offering detailed insights into abnormal areas.

Endoscopy produces internal images or videos of the gastrointestinal tract's walls or tissues, aiding in early detection and treatment, which significantly improves survival rates. An expert doctor is required to evaluate the findings, as abnormal regions may appear similar to normal ones and can occur at any location along the tract. Additionally, the doctor should be capable of making immediate diagnoses, especially with the advancements in modern endoscopic technology. As a result, computer-assisted techniques have become increasingly important and advantageous. Automated algorithms powered by computers are now used in clinical tasks to analyze images and videos for diagnosis, treatment planning, and prognosis. Deep learning, in particular, has emerged as a highly effective method for detection tasks. With the continuous advancement of hardware computing capabilities, its application is growing in popularity.

Despite considerable progress, achieving accurate and real-time automatic polyp detection remains a significant challenge. This is primarily due to the wide variations in polyp characteristics, such as shape, texture, size, and color, as well as the presence of artifacts that are similar to polyps during endoscopy. In this study, we propose a novel Gaussian Enhanced Euclidean norm Ghost attention (GEEG) module to enable reliable real-time polyp detection in endoscopic images and videos. This innovative attention mechanism enhances the features generated by the Ghost convolution’s cheap operations by improving the ability in the extraction of inter-channel and spatial information within the convolutional layers. The proposed module is incorporated into the YOLOv8 backbone, creating a new model named GEEG-YOLOv8, aimed at overcoming the obstacles in polyp

detection. Experimental evaluations on three public datasets reveal that our method surpasses current state-of-the-art approaches in both accuracy and detection speed.

# CHAPTER 1 INTRODUCTION

Colorectal cancer (CRC) ranks as the third most common cause of cancer-related deaths. While the five-year survival rate for colon cancer is around 68%, the corresponding rate for stomach cancer is just 44% [1]. Reducing CRC-related mortality can be greatly enhanced by early detection and removal of precancerous lesions, such as colon polyps, which may later progress to CRC. This chapter focuses on colorectal abnormalities, including both precancerous and cancerous conditions, and emphasizes the importance of early diagnosis. Additionally, problem statements, motivation, objectives, and contributions of the thesis will be presented.

## 1.1 Overview and Problem Statement

Around the world, nearly 5 million people are currently living with CRC at various stages and undergoing different treatments. In 2018, CRC accounted for 1.8 million new diagnoses and 881,000 deaths, representing 1 in 10 cancer-related fatalities [2]. The highest rates of CRC incidence are found in Europe, Australia, North America, and East Asia, while developing countries have lower rates, although these are on the rise [2]. Tracking incidence and mortality trends is complex, but Arnold et al. [3] identified three global patterns tied to countries development levels: increasing incidence and mortality in China, Russia, and Brazil; rising incidence but declining mortality in Canada, United Kingdom, Denmark, and Singapore; and reductions in both incidence and mortality in countries like the United States, Japan, and France [2], [4].

The increase in CRC cases is thought to be driven by factors such as diet, obesity, and lifestyle risks, while improved cancer treatment and management practices in developed nations are linked to lower mortality rates [2], [4]. Additionally, screening and early detection programs launched in the 1990s in United States and Japan are believed to have significantly improved survival rates and reduced mortality from CRC in these regions [2].

Globally, mortality rates from colorectal cancer (CRC) are rising rapidly and are predicted to continue increasing, particularly as life expectancy grows and the disease is more prevalent in older populations. Since 2012, the number of deaths caused by CRC has risen from 668,000 to 881,000 [2], [4]. The progression of CRC is a lengthy, multi-step process that evolves from precancerous lesions into malignant tumors, with

mortality rates strongly linked to the stage at which the disease is diagnosed [24]. Illustrative examples of several types of polyps are presented in Fig. 1.1.

The five-year survival rate is 90% for cases detected at a localized stage, 70% for regional cases, and only 10% for metastatic disease [4], [5]. These figures emphasize the urgent need for effective screening programs that can identify CRC in its early stages and detect precancerous lesions that can be removed to reduce cancer risk. Early detection and removal of precancerous polyps can dramatically improve survival rates, thus the importance of screening and detection as life-saving measures [4].

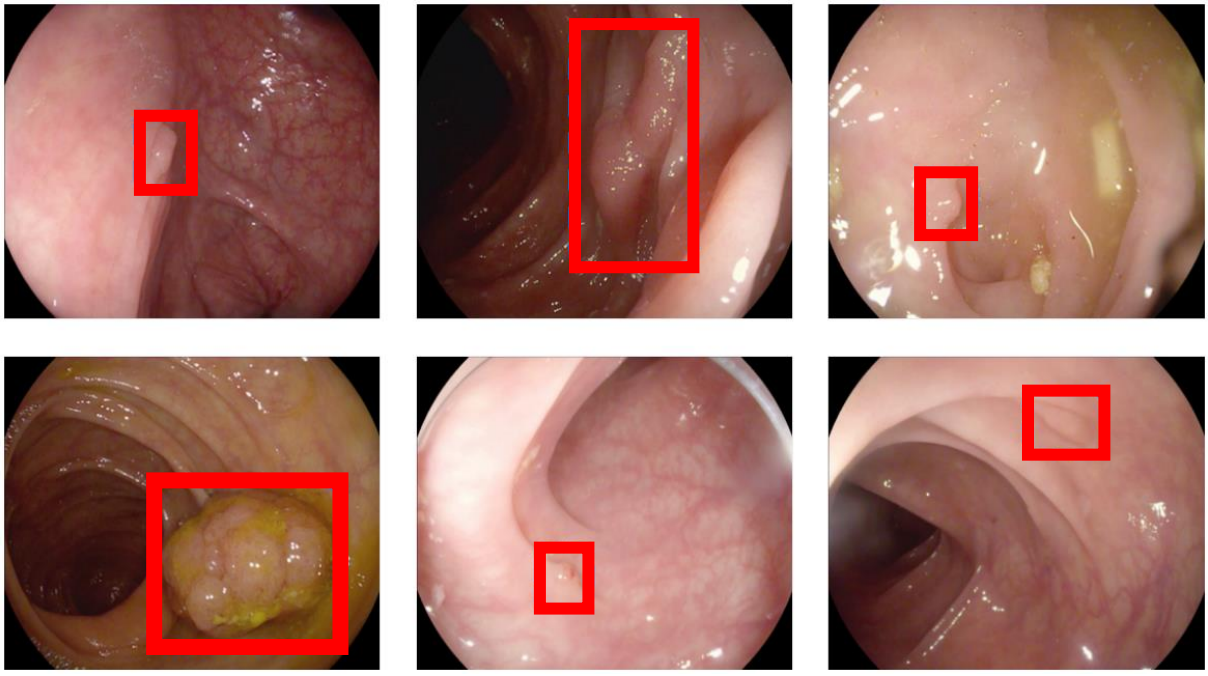


Fig. 1.1. Examples of neoplastic lesions in the colon tract [6].

A polyp is an abnormal growth that occurs on the inner lining of the colon. Polyps can differ in size, shape, attachment type to the colon wall, location, and histopathology. Most polyps do not cause clinical issues, only about 5% may develop into cancer over time. However, predicting the future behavior of a polyp based solely on its appearance is not possible. As a result, the majority of polyps are typically removed as a precaution [7]. Adenomatous and hyperplastic polyps are the most commonly detected types during screening colonoscopy. Although all adenomas have the potential to become cancerous, the majority are benign when detected. On the other hand, hyperplastic, mucosal, inflammatory, and hamartomatous polyps are considered non-cancerous and do not have malignant potential [8].

Globally, the prevalence of adenomas varies by region and correlates with the regional incidence rates of colorectal cancer. Autopsy studies from various regions have reported prevalence rates ranging from 22% to 61% [8]. Colonoscopy studies have indicated rates between 25% and 41% [8]. The risk of adenomas increases with age. Most of adenomas are less than 1.0 cm in diameter. A national study found that 38% of adenomas were 0.5 cm or smaller, 36% were between 0.6 and 1.0 cm, and 26% were larger than 1.0 cm. Of these, around 60% of patients had a single adenoma, while 40% had multiple adenomas [8]. Hyperplastic polyps, the most common type of non-neoplastic colon polyps, have a reported prevalence of 20% to 34% [8]. These polyps are typically small, that are 0.5 cm or less, and appear flat or slightly convex, often pale or similar in color to the surrounding mucosa [8].

The majority of colon polyps are less than 5 mm in diameter and sessile. Small to medium-sized polyps (6 to 9 mm in diameter) make up about 80% of all colon polyps. The prevalence of adenomas varies based on factors such as genetic risk, age, gender, obesity, and smoking habits. During a western screening/surveillance population, adenoma prevalence reaches approximately 50% [7]. Therefore, the demand for skilled physicians who can effectively distinguish between normal and abnormal mucosa is urgent. However, the limited number of such experts often results in doctors managing multiple patients simultaneously, especially in lower-level healthcare facilities. To address this, younger endoscopists are increasingly involved in the diagnostic process. These young physicians need thorough training from practical examples to enhance their abilities.

For the above reasons, computer-aided detection methods assist physicians in making more accurate diagnoses by serving as a second opinion. These systems can reduce the subjectivity involved in the diagnostic process and ease the burden on patients during regular follow-ups. Furthermore, they function as a training resource for junior doctors to learn how to recognize abnormal regions and gain experience with endoscopic tools operation.

## 1.2 Motivation

In recent years, computer-based technology for medical diagnosis and treatment has become a rapidly advancing field. Automated computer systems can assist

physicians by extracting clinically significant information from endoscopic images, serving as a valuable second opinion [9], [10], [11]. These systems are designed to detect and classify various disorders at different stages, enhancing survival rates. Automated systems typically involve multiple steps. Endoscopic videos or images are first preprocessed to improve quality by reducing noise or enhancing specific features. Abnormal regions are then automatically detected and, in some cases, segmented. In literature, computer-based models can either execute a single task (such as preprocessing, detection, segmentation, or classification) or perform multiple tasks simultaneously.

Deep learning methods have significantly improved the accuracy of automatic polyp detection in endoscopic videos and images. However, several challenges limit the performance of these methods. Polyp characteristics, such as shape, texture, and size, vary greatly, making it difficult for deep learning models to accurately identify them. Additionally, polyps can be obscured by factors like water flow, artifacts such as bubbles, light scatters during endoscopy procedure, or bodily tissues. Under specific camera view, polyps may appear very similar to the intestine wall. Unlike common moving object detection with stationary cameras, endoscopy involves a moving camera. The complex camera motion introduces unavoidable noises, such as motion blur, occlusion, and variations in brightness, which can affect detection accuracy. As a result, deep learning models may fail to detect polyps in certain images or videos.

In recent years, several studies have been proposed to enhance the performance of automatic polyp detection [12], [13], [14], [15], [16], [17], [18]. Many of these approaches are video-based object detection, utilizing features from extra dimensions to improve detection accuracy. However, this leads to slower detection speeds, making them unsuitable for practical applications. To address these issues, this thesis proposes a new deep learning framework for reliable real-time polyp detection on endoscopic images and videos. Unlike previous methods, it does not rely on extracting features from extra dimensions but still has high detection accuracy, ensuring a more efficient and practical solution.

## 1.3 Objectives and Contributions

The goal of this study is to develop an automatic processing approach that can detect polyps of various sizes for better CRC screening. The proposed method is required to locate polyps under different obstacles and artifacts while keeping real-time performance, reducing miss-detection rate during examination.

Our main contributions are as follows:

- A novel Gaussian Enhanced Euclidean norm attention mechanism is proposed to enhance the Ghost convolution's ability to extract inter-channel and spatial information.
- Gaussian Enhanced Euclidean norm Ghost attention module is incorporated into the backbone of YOLOv8 model, reducing the number of parameters and FLOPs while maintaining high detection accuracy, called GEEG-YOLOv8.
- Extensive experiments on three public datasets demonstrate superior performance compared to previous state-of-the-art methods.

# CHAPTER 2 COMPUTER-AIDED DIAGNOSIS FOR ENDOSCOPY

The use of computer-aided design (CAD) in diagnosing gastrointestinal abnormalities has significantly developed over the past decades. The purpose of this examination through imaging is to identify abnormal areas and classify their stages. Such advancements are beneficial in various clinical processes, including diagnosis, treatment planning, surgical preparation, and radiation therapy.

Endoscopic devices are employed to visualize the interior of a patient's body and obtain biopsy samples for organ examination. These devices come in a variety of designs to serve different purposes. Endoscopy offers numerous benefits, such as enabling doctors to investigate symptoms, locating abnormal regions, and facilitating surgical interventions.

This chapter provides an overview of different types of endoscopic tools and their applications in gastrointestinal examinations. Additionally, it includes a review of technical literature addressing various detection challenges and a detailed analysis of research studies that have implemented these techniques to enhance model performance.

## 2.1 Endoscopy Tools

Endoscopy is a non-surgical procedure used to examine various internal cavities within the human body [19]. There are different types of endoscopic techniques in medical field categorized based on the area being inspected, such as colonoscopy (colon), thoracoscopy (lungs), neuroendoscopy (brain and spine), etc. For examining the upper gastrointestinal tract where the esophagus is located, the procedure is known as esophagoscopy or gastroscopy [20]. During the procedure, a doctor inserts an endoscope (a flexible tube equipped with a camera and light attached to it) through the mouth into the esophagus, allowing visualization of the organs on a TV screen.

For examining the lower gastrointestinal tract, colonoscopy is an endoscopic procedure performed to investigate the large intestine and the distal part of the small intestine. This involves inserting a camera, light source, and additional instruments to examine the inner lining. Colonoscopy provides a real-time view of the colon's interior surface of the colon, enabling the collection of biopsies and the execution of therapeutic interventions for early stage neoplastic lesions [21].

The primary objective of colonoscopy and both the insertion and withdrawal phases of the endoscope are to traverse the zigzagged colon safely and efficiently while thoroughly examining the mucosal lining. During insertion, most colonoscopists concentrate on the technical aspects required to maneuver through the intestines, whereas during withdrawal, the focus shifts to inspecting the colon's surface [7]. Insertion can be difficult when the colon is highly serpentine and challenging to guide the scope through, whereas the withdrawal process depends on the doctor's ability to observe, the quality of bowel preparation, the nature of existing polyps, and their visibility [7].

To enhance the visualization of abnormal tissues during colonoscopy, various optical imaging technologies are available. Two of the most commonly utilized techniques are near-focus imaging and narrow-band imaging (NBI) [24]. Near-focus imaging allows the operator to closely approach the mucosa, providing high-resolution, magnified views of tissues and capillary networks. This is achieved by optimizing the lens structure integrated into the distal end of the colonoscope, enabling endoscopists to capture detailed mucosal surface information that cannot be achieved with electronic magnification alone [24]. NBI is often compared to traditional chromoendoscopy as it delivers enhanced contrast without the need for dyes. This technology operates by activating two electronic filters in the white light pathway, restricting the light to central wavelengths of 415 nm (blue) and 540 nm (green). These wavelengths coincide with the absorption peaks of hemoglobin, causing structures such as capillaries and veins to appear darker, thereby enhancing contrast against the surrounding mucosa [24]. By emphasizing surface microvasculature and the boundaries between different tissue types, NBI aids in the detailed assessment of gastrointestinal lesions, particularly neoplasia. It has also been applied to detect esophagitis, Barrett's esophagus, pit patterns in colon polyps and tumors, as well as identify dysplastic tissue in patients with ulcerative colitis [24].

## 2.2 Computer-aided Detection

Automatic detection of polyps in colonoscopy images and videos has been a highly researched topic over the last two decades. Initially, most of detection systems relied on simple and computationally efficient visual features like edges and color. However, as computational capabilities improved and data availability expanded, the focus moved

toward deep learning-based models [25]. This improvement in algorithms reflects the overall advancements in computer vision observed over the past decade.

### 2.2.1 Handcrafted Features

Early methods for polyp detection often relied on low-level shape descriptors, such as edge detectors [26], [27], [28], to approximate polyp boundaries. More sophisticated shape descriptors, such as Hessian filters and histograms of oriented gradients [29], were also employed to identify blob-like structures. Bernal et al. [30] introduced an innovative boundary model by detecting intensity valleys that typically surround a polyp. Their approach improved robustness against blood vessels and reflective highlights by incorporating metrics like completeness, continuity, and concavity. However, despite performing well on large datasets, these methods faced challenges in detecting small and flat polyps.

Other techniques utilized color and texture as key features for polyp detection [31], [32]. Wavelet transformations [33], [34] were applied to extract statistical texture features for distinguishing between image regions. Additionally, MPEG-7 shape and texture descriptors were used to detect polyps in capsule endoscopic images [35]. More simple descriptors such as local binary patterns [36] and co-occurrence matrices [37] were also investigated. Recently, Tajbakhsh et al. [38] proposed a hybrid approach combining shape and context by extracting image patches around edges and using a two-step classification process to filter out non-polyp patches.

Several handcrafted feature-based methods also utilized supervised learning techniques, such as linear discriminant analysis [33], support vector machines [29], [34], [37], [39], and random forests [38], to create their final classifiers. More recently, research has increasingly shifted toward end-to-end approaches.

### 2.2.2 Deep Learning

Driven by large-scale challenges like ImageNet [40], deep learning has transformed numerous domains in computer vision, outperforming traditional techniques in tasks such as classification, segmentation, detection, and tracking. In the context of polyp detection, deep learning developments were significantly enabled by the MICCAI endoscopic vision challenge [25], which provided the first substantial dataset suitable for training advanced deep learning models. This challenge also

introduced a validation framework that facilitated more effective comparisons between different approaches. Since then, additional datasets, such as Kvasir [41] and Nerthus [42], featuring diverse anatomical landmarks and pathological findings have been developed and released.

In recent years, many studies have been proposed to enhance the effectiveness of automatic polyp detection [12], [13], [14], [15], [16], [17], [18]. Most of these approaches focus on video-based object detection, leveraging features from the temporal dimension. Methods like RYCO [12] and AIPDT [13] integrated temporal information through discriminate correlation filter-based trackers. However, these techniques rely on accurate detection of the polyp in the initial frame, which is challenging due to image noise. Zheng et al. [14] refined detection results using optical flow, but significant variations between consecutive frames caused by complex camera movements reduce performance efficiency. STFT [15] introduced the Spatial-Temporal Feature Transformation to align features and minimize inconsistencies across multiple frames. While this method improves detection accuracy, its high computational cost limits its feasibility for real-time applications. YONA [16] enhanced detection accuracy and speed by extracting information from two consecutive frames employing the presented foreground temporal alignment, background dynamic alignment, and cross-frame box-assisted contrastive learning module.

For image-based object detection approaches, Shin et al. [17] utilized a region-based deep convolutional neural network combined with post-learning techniques to lower the false positive rate in polyp detection. Despite its effectiveness, this approach requires substantial computational resources, making real-time execution impractical. On the other hand, Wan et al. [18] incorporated an attention mechanism into the YOLOv5 [43] model to enable accurate and real-time polyp detection. However, the evaluation was conducted on a private dataset and a small public dataset, limiting the generalizability of the results.

In this thesis, we propose an image-based polyp detection framework that does not require feature extraction from extra dimensions like video-based object detection approaches but still has high detection accuracy and ensures real-time performance for practical applications.

# CHAPTER 3 PROPOSED METHOD

In this chapter, the proposed lightweight GEEG-YOLOv8 framework for polyp detection will be described. Then we will introduce the new Gaussian Enhanced Euclidean norm Ghost attention (GEEG) module which mitigates the weakness of Ghost Convolution (GhostConv) [44] in capturing global channel and spatial information.

## 3.1 GEEG-YOLOv8

The overall framework of GEEG-YOLOv8 is shown in Fig. 3.1.

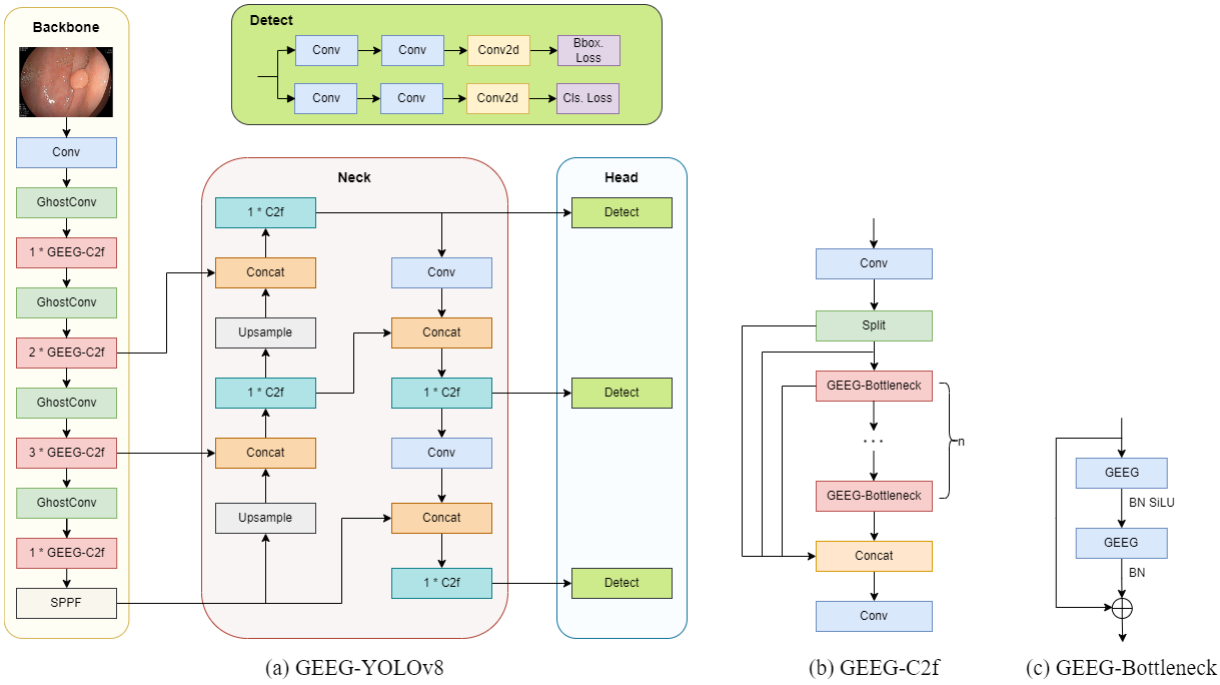



Fig. 3.1. The overall architecture of the proposed GEEG-YOLOv8. (Reproduced from our previous work Ref [1].)

The architecture is composed of a backbone, a neck, and a detection head. The backbone integrates several regular GhostConv [44] and GEEG-C2f modules. While the GEEG module improves information extraction capability, it also adds complexity to the model. To address this, GEEG is only incorporated into C2f to minimize the addition of excessive parameters. The GEEG-Bottleneck, illustrated in Fig. 3.1c, consists of two GEEG modules. The first GEEG module acts as a squeezing layer, reducing the input channels by half, while the second GEEG module expands the channels to match the shortcut path. Then the inputs and outputs of these two GEEG modules are connected by a residual connection. The first GEEG module is followed by Batch Normalization (BN) [45] and the SiLU [46] activation function, while only BN is applied after the

second module. The GEEG-C2f, depicted in Fig 3.1b, includes $n$ GEEG-Bottleneck with two parallel gradient flow branches, allowing it to capture richer gradient flow information while reducing parameters. The neck utilizes PANet [47], which combines feature maps from lower layers with deeper layers to enhance information flow. GEEG-YOLOv8 adopts an anchor-free approach with a decoupled detection head. The first branch of the detection head outputs the loss for object bounding box, while the second branch outputs the loss for object classification.

### 3.2 Gaussian Enhanced Euclidean norm Ghost attention module

By incorporating GhostConv into the backbone of YOLOv8 [48], the model's number of parameters and floating-point operations (FLOPs) can be significantly reduced. However, relying solely on GhostConv does not lead to a sufficient improvement in polyp detection performance. The cheap operations in GhostConv, which are usually $3 \times 3$ depth-wise convolution, only capture spatial information from the inherent feature maps created by $1 \times 1$ point-wise convolution, overlooking global dependency. Additionally, depth-wise convolution fails to consider the correlation between channel information. As a result, these cheap operations repeatedly extract local information generated from the inherent feature maps, limiting performance enhancement.

To address this, the Gaussian Enhanced Euclidean norm (GEE) attention mechanism is introduced after the cheap operations to enhance GhostConv's ability to extract inter-channel and spatial information. This attention mechanism is inspired by Convolutional Block Attention Module (CBAM) [49] and Gaussian Context Transformer (GCT) [50]. It is based on the hypothesis that smaller attention activations are linked to global contextual features with larger absolute values [50]. Euclidean norm measures the magnitude of a vector or a matrix, larger Euclidean norm means more deviation from the vector or matrix to its origin. Hence, the Euclidean norm of channel or spatial dimension is used as input for a Gaussian function to refine the output features of the channel and spatial mechanism.

As illustrated in Fig. 3.2, the feature maps generated by GEEG's cheap operations highlight more essential information, i.e., polyp features (marked by red arrows), compared to the naive GhostConv. GEEG's feature maps capture more fine-grained

features, whereas one feature map generated by GhostConv contains noise information, i.e., light scatter, distracting the model learning procedure.

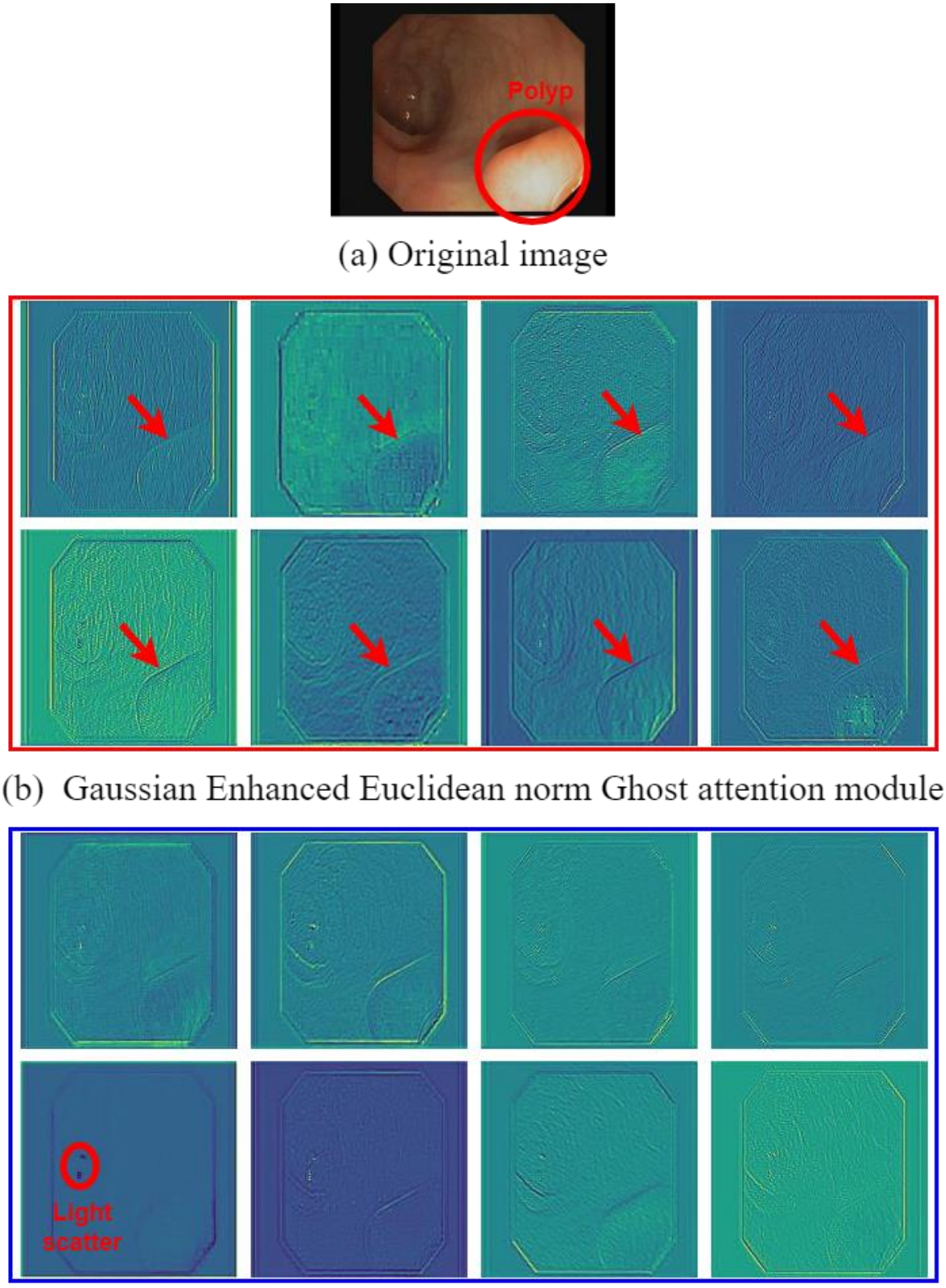


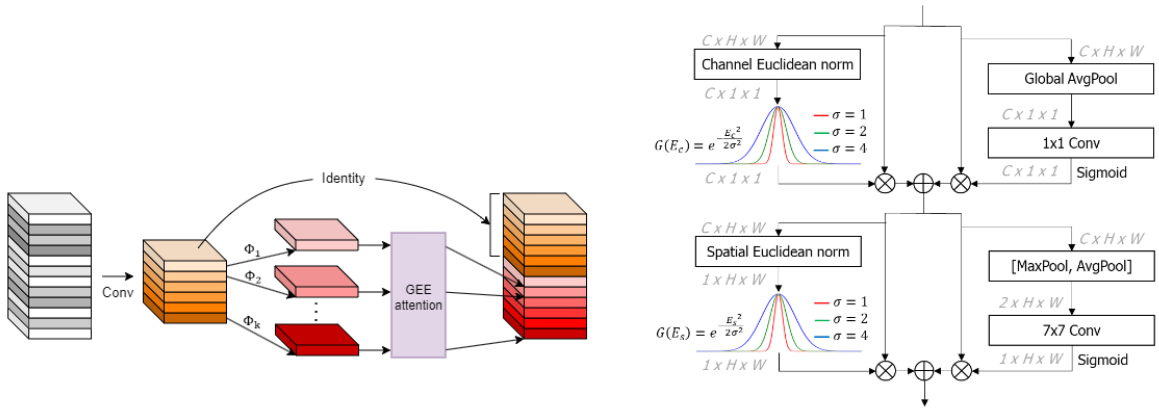

Fig. 3.2. Example of feature maps of an (a) original image generated by cheap operations in (b) Gaussian Enhanced Euclidean norm Ghost attention module and (c) Ghost convolution. (Reproduced from our previous work Ref [1].)

### 3.2.1 Gaussian Enhanced Euclidean norm attention

Self-attention [51] has been proven to be highly effective in capturing long-range global dependencies. However, its quadratic computational complexity makes it impractical for real-time deployment of hardware with limited resources. To address this challenge, we introduce the lightweight GEE attention mechanism, which is designed to enhance GhostConv performance with a negligible number of additional parameters. This approach is based on the hypothesis that smaller attention activations are linked to global contexts with larger absolute values. GEE consists of two modules: channel attention and spatial attention, each featuring two branches, as depicted in Fig. 3.3b.

The left branch of GEE differs from GCT in two main ways. First, instead of relying on global average pooling (GAP), we directly use the Euclidean norm of feature maps as the input for a Gaussian function. This is based on the intuition that the Euclidean norm represents the magnitude of a vector or matrix, with larger Euclidean norm indicating greater deviation from the origin. Constraining the Euclidean norm by Gaussian function enhances the model's generalization capabilities. Second, we extend this hypothesis to the spatial dimension as we argue that the above hypothesis is also true in spatial dimension, enabling the extraction of global spatial information.

The right branch is similar to CBAM but with a modification: in the channel attention module, only GAP is utilized, followed by a $1 \times 1$ convolution layer to decrease the number of model parameters.




(a) Gaussian Enhanced Euclidean norm Ghost attention module

(b) Gaussian Enhanced Euclidean norm attention

Fig. 3.3. Gaussian Enhanced Euclidean norm Ghost attention module block details. ⊗ represents broadcast element-wise multiplication and ⊕ denotes element-wise addition. E_c and E_s represent Channel Euclidean norm and Spatial Euclidean norm, respectively. (Reproduced from our previous work Ref [1].)

Concretely, given a feature map $\mathbf{F} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ as input, $C$ denotes the number of channel and $H$ and $W$ are spatial dimensions, GEE computes attention as follows.

$$\boldsymbol{F'} = \boldsymbol{M_c}(\boldsymbol{F}) \otimes \boldsymbol{F} \oplus \boldsymbol{G}(\boldsymbol{E_c}) \otimes \boldsymbol{F} \quad (1)$$

$$\boldsymbol{F''} = \boldsymbol{M_s}(\boldsymbol{F'}) \otimes \boldsymbol{F'} \oplus \boldsymbol{G}(\boldsymbol{E_s}) \otimes \boldsymbol{F'}, \quad (2)$$

where $\otimes$ represents broadcast element-wise multiplication and $\oplus$ denotes element-wise addition. $\mathbf{F''}$ is the final output. $\mathbf{E_c} \in \mathrm{R}^{\mathrm{C}\times1\times1}$ and $\mathbf{E_s} \in \mathrm{R}^{1\times\mathrm{H}\times\mathrm{W}}$ represents Channel Euclidean norm and Spatial Euclidean norm, respectively, and is formulated as follows.

$$\boldsymbol{E_c} = \left\{ e_{ck} = \sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H} \boldsymbol{F_k}(i,j)^2} : k \in \{1, \dots, C\} \right\} \quad (3)$$

$$\boldsymbol{E_s} = \left\{ e_{sij} = \sqrt{\sum_{k=1}^{C} \boldsymbol{F'_{ij}}(k)^2} : i \in \{1, \dots, W\}, j \in \{1, \dots, H\} \right\}. \quad (4)$$

A Gaussian function, expressed as $\mathbf{G}(\mathrm{x}) = \exp\left(-\frac{\mathrm{x}^2}{2\sigma^2}\right)$, uses $x$ as its input, with maximum value of 1, mean of 0, and standard deviation σ, to align with the hypothesis regarding the relationship between global contexts and attention activations. Increasing the standard deviation reduces the difference between each attention activation. The effect of standard deviation on model performance will be analyzed in the experimental section.

$\mathbf{M_c} \in \mathrm{R}^{\mathrm{C}\times1\times1}$ and $\mathbf{M_s} \in \mathrm{R}^{1\times\mathrm{H}\times\mathrm{W}}$ is channel attention map and spatial attention map, respectively, and computed as

$$\boldsymbol{M_c}(\boldsymbol{F}) = Sigmoid\left(f^{1\times1}\left(GAP(\boldsymbol{F})\right)\right), \quad (5)$$

$$\boldsymbol{M_s}(\boldsymbol{F}) = Sigmoid\left(f^{7\times7}([MaxPool(\boldsymbol{F});\ AvgPool(\boldsymbol{F})])\right), \quad (6)$$

where $f^{k\times k}$ represents a convolution operation with kernel size of k × k.

The left branch refines the right branch's output by putting more attention to potential concealed polyp features that possess low activation values while suppressing the importance of polyp-like noise features to the model.

### 3.2.2 Combine with Ghost Convolution

GEE attention mechanism is integrated after cheap operations in GhostConv to form GEEG module, as depicted in Fig. 3.3a. Although conventional GhostConv can reduce computational cost, its capacity for information extraction is limited, as the cheap

operations only capture local information from inherent feature maps produced by 1 × 1 point-wise convolution. By incorporating GEE attention mechanism, GhostConv's ability to capture long-range global channel and spatial information is significantly elevated. For an input feature map $\mathbf{F} \in \mathrm{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$, GEEG performs two steps. First, it generates the inherent feature map $\mathbf{Y}' \in \mathrm{R}^{\mathrm{C}'\times\mathrm{H}\times\mathrm{W}}$ by

$$\boldsymbol{Y}' = f^{1\times1}(\boldsymbol{F}). \tag{7}$$

Then the output feature map $\mathbf{Y} \in \mathrm{R}^{\mathrm{C_{out}}\times\mathrm{H}\times\mathrm{W}}$ is computed as follows.

$$\boldsymbol{Y} = Concat\left(\left[\boldsymbol{Y}', GEE\left(\Phi_{dp}(\boldsymbol{Y}')\right)\right]\right), \tag{8}$$

where $\Phi_{\mathrm{dp}}$ denotes depth-wise convolution operation, and $\mathrm{C}' < \mathrm{C_{out}}$.

# CHAPTER 4 EXPERIMENTS

In this section, the evaluation measures for the proposed framework are presented. Afterward, the dataset used, implementation details and evaluation protocols are described. Finally, comprehensive experimental results are presented and discussed in terms of quantitative and qualitative evaluations.

## 4.1 Datasets

The following datasets were utilized for our experiments: Kvasir-SEG [41], NeoPolyp-Small [52], PolypsSet [53], LDPolypVideo [54], CVC-ClinicDB [30], ETIS-LaribPolypDB [30], and SUN [55]. To assess the generalizability of the proposed method, we combined the first four datasets as the training set and evaluated the model on the remaining three datasets. Images with identical viewpoint and distance of the same polyp, as well as those that were excessively blurry or exhibited significant artifacts, were manually excluded from the training set since they did not provide valuable information for learning. This resulted in a total of 24,734 training images. For the test datasets, CVC-ClinicDB and ETIS-LaribPolypDB include 612 and 196 polyp images, respectively, while the SUN database contains 49,136 polyp frames derived from 100 distinct polyp video sequences.

## 4.2 Experimental Setup

All experiments were performed using an NVIDIA RTX 4070 GPU with 12GB of VRAM. The models were developed using the PyTorch framework. The optimization process employed the SGD optimizer with an initial learning rate of 0.01, a momentum of 0.937, and a weight decay rate of 0.0005. All images were resized to $640 \times 640$, with a batch size of 32. The models were trained for 400 epochs. Detection performance metrics included precision (P), recall (R), mean average precision at an IoU threshold of 0.5 (mAP@50), and mean average precision across IoU thresholds from 0.5 to 0.95 (mAP@50-95). Models' complexity and speed were measured in terms of the number of parameters, GFLOPs, and frames per second (FPS).

## 4.3 Experimental Results

### 4.3.1 Impact of Standard Deviation on Model Performance

This section examines the effect of the standard deviation σ in the Gaussian function $\mathbf{G}(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$ on the detection performance of GEEG-YOLOv8. The results are illustrated in Table 4.1. It can be observed that as σ increases, the model performance initially improves but then gradually declines. The optimal performance is achieved when σ is set to 4. This behavior is logical, as when variance that is too large diminishes the differences in attention activations across channel and spatial dimension, making it harder to effectively suppress global noises. On the other hand, a variance that is too small can overly restrict the significance of other important features and inadvertently emphasize noise contexts.

Table 4.1 Detection results of GEEG-YOLOv8 on SUN dataset with different standard deviation σ

| $\sigma$ | 1 | 2 | 4 | 6 |
|---|---|---|---|---|
| P | 84.86 | 86.09 | 87.23 | 87.12 |
| R | 69.26 | 69.83 | 71.56 | 67.27 |
| mAP@50 | 80.51 | 81.21 | 82.49 | 80.26 |
| mAP@50-95 | 44.06 | 45.04 | 45.41 | 44.28 |

### 4.3.2 Comparisons with State-of-the-arts Methods

The quantitative comparison of the proposed GEEG-YOLOv8 model with other detection models is shown in Table 4.2. Overall, our model outperforms other methods across all metrics on the SUN and CVC-ClinicDB datasets. For the ETIS-LaribPolypDB dataset, GEEG-YOLOv8 achieves the highest scores in mAP@50 and mAP@50-95, while securing the second-best results for precision (P) and recall (R). Compared to the runner-up models, GEEG-YOLOv8 delivers notable improvements, including a 3% increase in P and mAP@50-95 on SUN, a 0.5% improvement in R and mAP@50 on CVC-ClinicDB, and a 2% gain in mAP@50 and mAP@50-95 on ETIS-LaribPolypDB. The proposed GEEG-YOLOv8 also outperforms other models specifically designed for polyp detection. Model complexity and FPS rate are illustrated in Table 4.3. Thanks to GEEG's efficient and lightweight design, the number of parameters and GFLOPs in our model are reduced compared to the original YOLOv8. While it does not achieve the lowest number of parameters and GFLOPs, GEEG-YOLOv8 obtains the second-highest FPS at 303. Our proposed method achieves new state-of-the-art in polyp detection

accuracy with low model complexity, showing its potential for real-world applications where real-time processing is essential.

Table 4.2 Detection performance comparisons with other models (The best score is denoted as red, while the runner-up score is denoted as blue)

| Method | SUN | | | | CVC-ClinicDB | | | | ETIS-LaribPolypDB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | mAP@50 | mAP@50-95 | P | R | mAP@50 | mAP@50-95 | P | R | mAP@50 | mAP@50-95 |
| STFT [15] | 81.50 | 71.45 | 80.69 | 40.12 | - | - | - | - | - | - | - | - |
| AIPDT [13] | 80.37 | 69.78 | 79.21 | 38.55 | - | - | - | - | - | - | - | - |
| YONA [16] | 83.30 | 71.52 | 81.43 | 41.89 | - | - | - | - | - | - | - | - |
| Wan et al [18] | 82.81 | 70.39 | 80.11 | 40.07 | 82.83 | 73.22 | 76.93 | 49.80 | 73.95 | 76.70 | 80.34 | 60.88 |
| EfficientDet-D0 [56] | 74.86 | 58.23 | 63.02 | 27.57 | 77.19 | 73.15 | 75.70 | 47.33 | 70.10 | 73.78 | 75.51 | 50.26 |
| YOLOv3-tiny [57] | 74.70 | 57.01 | 62.69 | 26.96 | 76.04 | 72.62 | 74.68 | 46.80 | 71.29 | 75.00 | 74.72 | 48.68 |
| YOLOv6s [58] | 84.58 | 68.25 | 74.94 | 37.50 | 85.42 | 74.12 | 79.12 | 51.48 | 89.94 | 77.55 | 82.66 | 61.69 |
| YOLOv7-tiny [59] | 84.21 | 71.52 | 81.78 | 41.95 | 79.59 | 74.10 | 78.84 | 48.82 | 86.66 | 75.53 | 81.17 | 59.40 |
| YOLOv8s [48] | 83.76 | 63.95 | 76.14 | 40.71 | 82.48 | 71.24 | 81.03 | 55.24 | 85.13 | 73.01 | 81.77 | 60.92 |
| **GEEG-YOLOv8 (ours)** | 87.23 | 71.56 | 82.49 | 45.41 | 85.48 | 74.67 | 81.54 | 55.29 | 87.59 | 76.02 | 84.19 | 63.56 |

Table 4.3 Model complexity and frames per second rate comparisons

| Method | Params | GFLOPs | FPS |
|---|---|---|---|
| STFT [15] | - | - | 12.5 |
| AIPDT [13] | - | - | 72 |
| YONA [16] | - | - | 46.3 |
| Wan et al [18] | - | - | 45 |
| EfficientDet-D0 [56] | 3.9M | 2.4 | 209 |
| YOLOv3-tiny [57] | 8.6M | 12.9 | 370 |
| YOLOv6s [58] | 18.5M | 45.2 | 275 |
| YOLOv7-tiny [59] | 6.0M | 13.0 | 208 |
| YOLOv8s [48] | 11.1M | 28.4 | 300 |
| **GEEG-YOLOv8 (ours)** | 8.5M | 21.2 | 303 |

The qualitative comparisons of GEEG-YOLOv8 with other models across three datasets in Fig. 4.1. The proposed approach can effectively detect polyps in different scenarios, such as blurring effects in the SUN dataset. It is also capable of identifying small and flat polyps in the CVC-ClinicDB and ETIS-LaribPolypDB dataset.

Furthermore, GEEG-YOLOv8 shows a robust performance in distinguishing polyps that look similar to the intestinal wall.

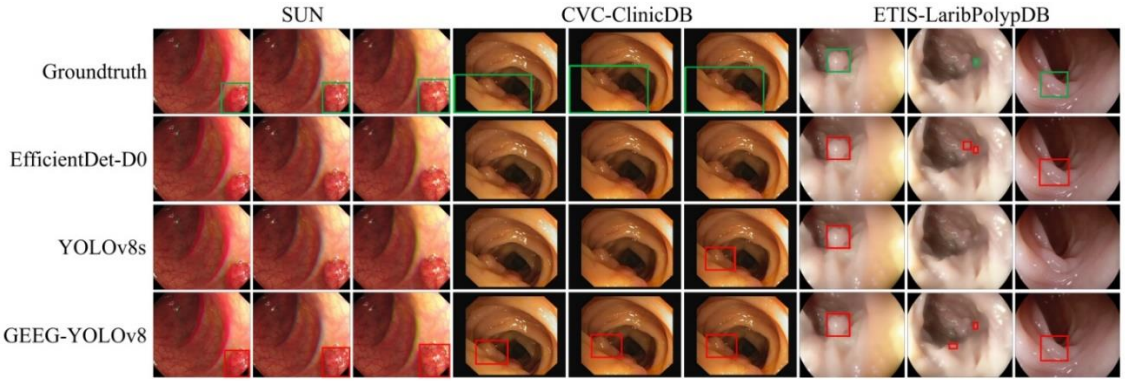



Fig. 4.1. Qualitative comparisons of polyp detection on three datasets. Green and red denote the ground truth and model prediction, respectively. (Reproduced from our previous work Ref [1].)

### 4.3.3 Ablation Studies

To evaluate the impact of individual components in the proposed method, ablation experiments were performed on the SUN dataset, with results presented in Table 4.4. In method (a), only GhostConv was used in the backbone of the original YOLOv8. While this approach achieved the highest FPS, its detection performance was sub-optimal due to the limitation of GhostConv ability in capturing global information. Method (b) added the channel and spatial attention mechanism (the right branch in Fig. 3.3b) after GhostConv's cheap operations, leading to a 3% improvement in R, mAP@50, and mAP@50-95, with only an additional 0.1M parameters. Method (c) utilized only GEE (the left branch in Fig. 3.3b) to GhostConv and achieved approximately a 1% gain in all four detection metrics compared to method (b), without adding extra parameters. By combining all three improvements in method (d), the highest polyp detection performance was achieved without increasing parameters and GFLOPs compared to method (b), with only a slight reduction in FPS. Method (d) achieved P, R, mAP@50, and mAP@50-95 scores of 87.23%, 71.56%, 82.49%, and 45.41%, respectively.

Table 4.4 Ablation results on SUN dataset of the proposed method

| Method | GhostConv | $\boldsymbol{M}(x)$ | $\boldsymbol{G}(x)$ | P | R | mAP@50 | mAP@50-95 | Params | GFLOPs | FPS |
|---|---|---|---|---|---|---|---|---|---|---|
| (a) | ✓ | | | 85.05 | 64.51 | 77.25 | 41.95 | 8.4M | 21.2 | 322 |
| (b) | ✓ | ✓ | | 85.29 | 67.22 | 80.10 | 44.21 | 8.5M | 21.2 | 310 |
| (c) | ✓ | | ✓ | 86.09 | 69.32 | 81.11 | 44.85 | 8.4M | 21.2 | 312 |
| (d) | ✓ | ✓ | ✓ | 87.23 | 71.56 | 82.49 | 45.41 | 8.5M | 21.2 | 303 |

# CHAPTER 5 CONCLUSION AND FUTURE WORKS

This thesis proposes a novel Gaussian Enhanced Euclidean norm Ghost attention module designed to enhance the accuracy and speed of polyp detection in endoscopic images and videos. The module utilizes a Gaussian function applied to the Euclidean norm of channel and spatial dimension, refining the output features of channel and spatial attention mechanism. This approach improves Ghost convolution's ability to capture long-range global contextual information. The module is incorporated into the backbone of YOLOv8, forming a new model called GEEG-YOLOv8. Extensive experimental evaluations show that the proposed method achieves robust generalization without requiring information from additional dimensions. With a minimal increase in model parameters and GFLOPs, GEEG-YOLOv8 achieves state-of-the-art performance in polyp detection across three datasets. Future work will focus on refining the neck and detection head to further boost the model's detection capabilities.

## REFERENCES


[1] J. Asplund, J. H. Kauppila, F. Mattsson, and J. Lagergren, "Survival Trends in Gastric Adenocarcinoma: A Population-Based Study in Sweden," *Ann. Surg. Oncol.*, vol. 25, no. 9, pp. 2693–2702, Sep. 2018, doi: 10.1245/s10434-018-6627-y.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018, doi: 10.3322/caac.21492.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global patterns and trends in colorectal cancer incidence and mortality," *Gut*, vol. 66, no. 4, pp. 683–691, Apr. 2017, doi: 10.1136/gutjnl-2015-310912.

[4] J. H. Scholefield and C. Eng, *Colorectal Cancer: Diagnosis and Clinical Management*. John Wiley & Sons, 2014.

[5] A. B. Benson (III), A. B. Chakravarthy, S. R. Hamilton, S. H. MD, and E. R. Sigurdson, *Cancers of the Colon and Rectum: A Multidisciplinary Approach to Diagnosis and Management*. Demos Medical Publishing, 2013.

[6] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 9, no. 2, pp. 283–293, Mar. 2014, doi: 10.1007/s11548-013-0926-3.

[7] J. D. Waye, J. Aisenberg, and P. H. Rubin, *Practical Colonoscopy*. John Wiley & Sons, 2013.

[8] A. J. Markowitz and S. J. Winawer, "Management of colorectal polyps," *CA. Cancer J. Clin.*, vol. 47, no. 2, pp. 93–112, 1997, doi: 10.3322/canjclin.47.2.93.

[9] P. S. Hiremath, B. V. Dhandra, I. Humnabad, R. Hegadi, and G. G. Rajput, "Detection of esophageal Cancer (Necrosis) in the Endoscopic images using color image segmentation," *National Conference on Document Analysis and Recognition (NCDAR 2003), Mandya, India*, pp. 417–422, Jul. 2003.

[10] J. de Groof *et al.*, "The Argos project: The development of a computer-aided detection system to improve detection of Barrett's neoplasia on white light

endoscopy," *United Eur. Gastroenterol. J.*, vol. 7, no. 4, pp. 538–547, May 2019, doi: 10.1177/2050640619837443.

[11]Y.-Y. Zhao *et al.*, "Computer-assisted diagnosis of early esophageal squamous cell carcinoma using narrow-band imaging magnifying endoscopy," *Endoscopy*, vol. 51, no. 04, pp. 333–341, Apr. 2019, doi: 10.1055/a-0756-8754.

[12]R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern Recognit.*, vol. 83, pp. 209–219, Nov. 2018, doi: 10.1016/j.patcog.2018.05.026.

[13]Z. Zhang *et al.*, "Asynchronous in Parallel Detection and Tracking (AIPDT): Real-Time Robust Polyp Detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds., Cham: Springer International Publishing, pp. 722–731, 2020. doi: 10.1007/978-3-030-59716-0_69.

[14]H. Zheng, H. Chen, J. Huang, X. Li, X. Han, and J. Yao, "Polyp Tracking in Video Colonoscopy Using Optical Flow With an On-The-Fly Trained CNN," in *2019 IEEE International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 79–82. Apr. 2019. doi: 10.1109/ISBI.2019.8759180.

[15]L. Wu, Z. Hu, Y. Ji, P. Luo, and S. Zhang, "Multi-frame Collaboration for Effective Endoscopic Video Polyp Detection via Spatial-Temporal Feature Transformation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Springer International Publishing, pp. 302–312. Sep. 2021. doi: 10.1007/978-3-030-87240-3_29.

[16]Y. Jiang, Z. Zhang, R. Zhang, G. Li, S. Cui, and Z. Li, "YONA: You Only Need One Adjacent Reference-Frame for Accurate and Fast Video Polyp Detection," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, Springer Nature Switzerland, pp. 44–54. 2023. doi: 10.1007/978-3-031-43904-9_5.

[17]Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, "Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018, doi: 10.1109/ACCESS.2018.2856402.

[18]J. Wan, B. Chen, and Y. Yu, “Polyp Detection from Colorectum Images by Using Attentive YOLOv5,” *Diagnostics*, vol. 11, no. 12, Art. no. 12, Dec. 2021, doi: 10.3390/diagnostics11122264.

[19]T. Gotoda, “Endoscopic resection of early gastric cancer,” *Gastric Cancer*, vol. 10, no. 1, pp. 1–11, Feb. 2007, doi: 10.1007/s10120-006-0408-1.

[20]M. Liedlgruber and A. Uhl, “Computer-Aided Decision Support Systems for Endoscopy in the Gastrointestinal Tract: A Review,” *IEEE Rev. Biomed. Eng.*, vol. 4, pp. 73–88, 2011, doi: 10.1109/RBME.2011.2175445.

[21]D. K. Rex *et al.*, “Quality Indicators for Colonoscopy,” *Off. J. Am. Coll. Gastroenterol. ACG*, vol. 101, no. 4, p. 873, Apr. 2006.

[22]“Upper Endoscopy (Discharge Care).” Accessed: Jan. 13, 2025. [Online]. Available: https://www.drugs.com/cg/upper-endoscopy-discharge-care.html

[23]“Lower Gi Endoscopy Colonoscopy in kalyan | Gastroenterologist doctor,” The Gastro Clinic. Accessed: Jan. 13, 2025. [Online]. Available: https://www.gastroclinic.co/lower-gi-endosocpy-colonoscopy/

[24]Z. He, P. Wang, Y. Liang, Z. Fu, and X. Ye, “Clinically Available Optical Imaging Technologies in Endoscopic Lesion Detection: Current Status and Future Perspective,” *J. Healthc. Eng.*, vol. 2021, no. 1, p. 7594513, 2021, doi: 10.1155/2021/7594513.

[25]J. Bernal *et al.*, “Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results From the MICCAI 2015 Endoscopic Vision Challenge,” *IEEE Trans. Med. Imaging*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017, doi: 10.1109/TMI.2017.2664042.

[26]J. Kang and R. Doraiswami, “Real-time image processing system for endoscopic applications,” in *CCECE 2003 - Canadian Conference on Electrical and Computer Engineering. Toward a Caring and Humane Technology (Cat. No.03CH37436)*, pp. 1469–1472 vol.3, May 2003. doi: 10.1109/CCECE.2003.1226181.

[27]S. M. Krishnan, X. Yang, K. L. Chan, S. Kumar, and P. M. Y. Goh, “Intestinal abnormality detection from endoscopic images,” *IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, pp. 895–898 vol.2, Nov. 1998. doi: 10.1109/IEMBS.1998.745583.

[28]S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. de Groen, “Polyp Detection in Colonoscopy Video using Elliptical Shape Feature,” *IEEE International Conference on Image Processing*, pp.II-465-II–468, Sep. 2007. doi: 10.1109/ICIP.2007.4379193.

[29]Y. Iwahori, T.Shimohara, A. Hattori, R. J. Woodham, S. Fukui, M. K. Bhuyan, and K. Kasugai, “Automatic Polyp Detection in Endoscope Images Using a Hessian Filter.,” *International Conference on Machine Vision Applications (MVA2013)*, pp. 21–24, May 2013.

[30]J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Comput. Med. Imaging Graph.*, vol. 43, pp. 99–111, Jul. 2015, doi: 10.1016/j.compmedimag.2015.02.007.

[31]B. V. Dhandra, R. Hegadi, M. Hangarge, and V. S. Malemath, “Analysis of Abnormality in Endoscopic images using Combined HSI Color Space and Watershed Segmentation,” *International Conference on Pattern Recognition (ICPR’06)*, pp. 695–698. Aug. 2006. doi: 10.1109/ICPR.2006.268.

[32]M. P. Tjoa and S. M. Krishnan, “Feature extraction for the analysis of colon status from the endoscopic images,” *Biomed. Eng. OnLine*, vol. 2, no. 1, p. 9, Apr. 2003, doi: 10.1186/1475-925X-2-9.

[33]S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, “Computer-aided tumor detection in endoscopic video using color wavelet features,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003, doi: 10.1109/TITB.2003.813794.

[34]P. Li, K. L. Chan, and S. M. Krishnan, “Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, pp. 670–675, vol. 2, Jun. 2005. doi: 10.1109/CVPR.2005.201.

[35]M. T. Coimbra and J. P. S. Cunha, “MPEG-7 Visual Descriptors—Contributions for Automated Feature Extraction in Capsule Endoscopy,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 628–637, May 2006, doi: 10.1109/TCSVT.2006.873158.

[36]S. Gross, T. Stehle, A. Behrens, R. Auer, T. Aach, R. Winograd, C. Trautwein, and J. Tischendorf, “A comparison of blood vessel features and local binary patterns for colorectal polyp classification,” *Medical Imaging 2008: Computer-Aided Diagnosis*, SPIE, pp. 758–765, Apr. 2008. doi: 10.1117/12.810996.

[37]S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, “Texture-Based Polyp Detection in Colonoscopy,” in *Bildverarbeitung für die Medizin 2009*, H.-P. Meinzer, T. M. Deserno, H. Handels, and T. Tolxdorff, Eds., Berlin, Heidelberg: Springer, 2009, pp. 346–350. doi: 10.1007/978-3-540-93860-6_70.

[38]N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information,” *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 630–644, Feb. 2016, doi: 10.1109/TMI.2015.2487997.

[39]L. A. Alexandre, J. Casteleiro, and N. Nobreinst, “Polyp Detection in Endoscopic Video Using SVMs,” *Knowledge Discovery in Databases: PKDD 2007*, Springer, pp. 358–365, 2007. doi: 10.1007/978-3-540-74976-9_34.

[40]O. Russakovsky *et al.*, “ImageNet Large Scale Visual Recognition Challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[41]D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-SEG: A Segmented Polyp Dataset,” *MultiMedia Modeling*, Springer International Publishing, pp. 451–462, 2020. doi: 10.1007/978-3-030-37734-2_37.

[42]K. Pogorelov *et al.*, “Nerthus: A Bowel Preparation Quality Video Dataset,” *ACM on Multimedia Systems Conference 2017 (*MMSys’17). pp. 170–174, Jun. 2017. doi: 10.1145/3083187.3083216.

[43]G. Jocher, *YOLOv5 by Ultralytics*. (May 2020). Python. doi: 10.5281/zenodo.3908559.

[44]K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “GhostNet: More Features From Cheap Operations,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), pp. 1580–1589, Jun. 2020.

[45]S. Ioffe, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *ArXiv Prepr. ArXiv150203167*, 2015.

[46]S. Elfwing, E. Uchibe, and K. Doya, “Sigmoid-weighted linear units for neural network function approximation in reinforcement learning,” *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018, doi: 10.1016/j.neunet.2017.12.012.

[47]S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path Aggregation Network for Instance Segmentation,” IEEE Conference on Computer Vision and Pattern Recognition, (CVPR 2018), pp. 8759–8768, Jun. 2018.

[48]G. Jocher, J. Qiu, and A. Chaurasia, *Ultralytics YOLO*. (Jan. 2023). Python. Accessed: Jan. 13, 2025. [Online]. Available: https://github.com/ultralytics/ultralytics

[49]S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” European Conference on Computer Vision (ECCV 2018), pp. 3–19, Sep. 2018.

[50]D. Ruan, D. Wang, Y. Zheng, N. Zheng, and M. Zheng, “Gaussian Context Transformer,” IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR 2021), pp. 15129–15138, Jun. 2021.

[51]A. Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ArXiv Prepr. ArXiv201011929*, 2020.

[52]P. Ngoc Lan *et al.*, “NeoUNet : Towards Accurate Colon Polyp Segmentation and Neoplasm Detection,” *Advances in Visual Computing*, Springer International Publishing, pp. 15–28, 2021. doi: 10.1007/978-3-030-90436-4_2.

[53]K. Li *et al.*, “Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations,” *PLOS ONE*, vol. 16, no. 8, p. e0255809, Aug. 2021, doi: 10.1371/journal.pone.0255809.

[54]Y. Ma, X. Chen, K. Cheng, Y. Li, and B. Sun, “LDPolypVideo Benchmark: A Large-Scale Colonoscopy Video Dataset of Diverse Polyps,” *Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, Springer International Publishing, pp. 387–396, 2021. doi: 10.1007/978-3-030-87240-3_37.

[55]M. Misawa *et al.*, “Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video),” *Gastrointest. Endosc.*, vol. 93, no. 4, pp. 960-967.e3, Apr. 2021, doi: 10.1016/j.gie.2020.07.060.

[56]M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and Efficient Object Detection,” presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2020), pp. 10781–10790, Jun. 2020.

[57]A. Farhadi and J. Redmon, “Yolov3: An incremental improvement,” in *Computer vision and pattern recognition*, Springer, pp. 1–6, 2018.

[58]C. Li *et al.*, “YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications,” Sep. 2022, *arXiv*: arXiv:2209.02976. doi: 10.48550/arXiv.2209.02976.

[59]C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023), pp. 7464–7475, Jun. 2023.

## LIST OF PUBLICATIONS

[1] P. T. Nguyen and H. Watanabe, “GEEG-YOLOv8: Gaussian Enhanced Euclidean Norm Ghost Attention for Real-Time Polyp Detection,” IEEE International Conference on Image Processing (ICIP 2024), pp.3057-3063, Oct. 2024. DOI: 10.1109/ICIP51287.2024.10648065

[2] P. T. Nguyen and H. Watanabe, “A Real-time Polyp Detection Method Based on GhostAtt-YOLOv8,” IIEEJ International Conference on Image Electronics and Visual Computing (IEVC2024), 8B-3, Mar. 2024.