GEEG-YOLOV8: GAUSSIAN ENHANCED EUCLIDEAN NORM GHOST ATTENTION FOR REAL-TIME POLYP DETECTION

Phuong Thao Nguyen *Hiroshi Watanabe*

Graduate School of Fundamental Science and Engineering, Waseda University

2024 IEEE International Conference on Image Processing (ICIP) | 979-8-3503-4939-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICIP51287.2024.10648065

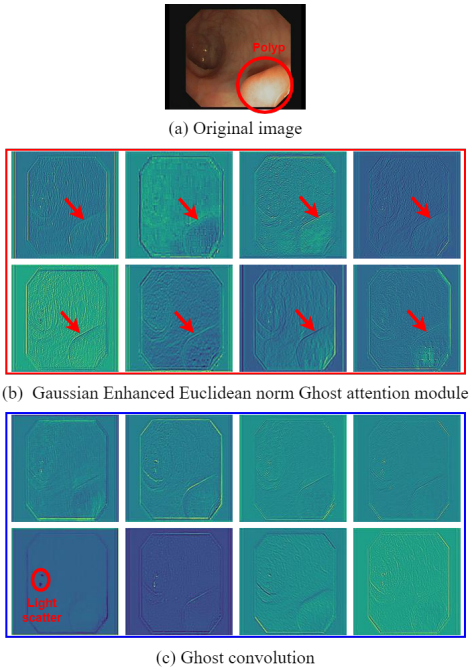
ABSTRACT

Research on computer-aided polyp detection in gastrointestinal endoscopy has spanned the past few decades. Despite notable progress, the challenge of achieving automatic accurate and real-time polyp detection remains unresolved. This is because of the large differences in polyp characteristics such as shape, texture, size, and color, and the artifacts that are similar to polyp during endoscopy procedure. In this paper, we propose a novel Gaussian Enhanced Euclidean norm Ghost attention (GEEG) module for reliable real-time polyp detection on endoscopic images and videos. The new attention mechanism strengthens the features generated by Ghost convolution's cheap operations by increasing the ability to extract inter-channel and spatial information inside the convolution layer. This module is integrated into the backbone of YOLOv8, creating a new model named GEEG-YOLOv8, to overcome above obstacles in polyp detection. Experiment results on three public datasets show that our proposed method outperforms existing state-of-the-art methods in both accuracy and speed.

***Index Terms*—** Attention mechanism, yolov8, deep learning, polyp detection, medical image analysis

1. INTRODUCTION

Colorectal cancer (CRC) stands as the third-leading cause of cancer-related death. While colon cancer boasts a five-year survival rate of around 68%, the corresponding rate for stomach cancer is merely 44% [1]. Addressing CRC-related mortality can be significantly advanced by identifying and eliminating precancerous lesions like colon polyps, which carry the potential to develop into CRC later on. Hence, early detection of polyps is crucial for enhancing survival rates. Colonoscopy is an invasive medical procedure which involves the use of a flexible endoscope by an endoscopist to examine and treat the colon. It is widely considered the optimal diagnostic tool for early detection and treatment of polyps, making it the preferred choice for gastroenterologists in screening for colon-related issues. However, about 25% of polyps are missed during inspection due to endoscopist exhaustion [1]. Therefore, computer-based detection methods come to the aid of physicians for a more accurate diagnosis and reduce miss-detection rate of polyps.



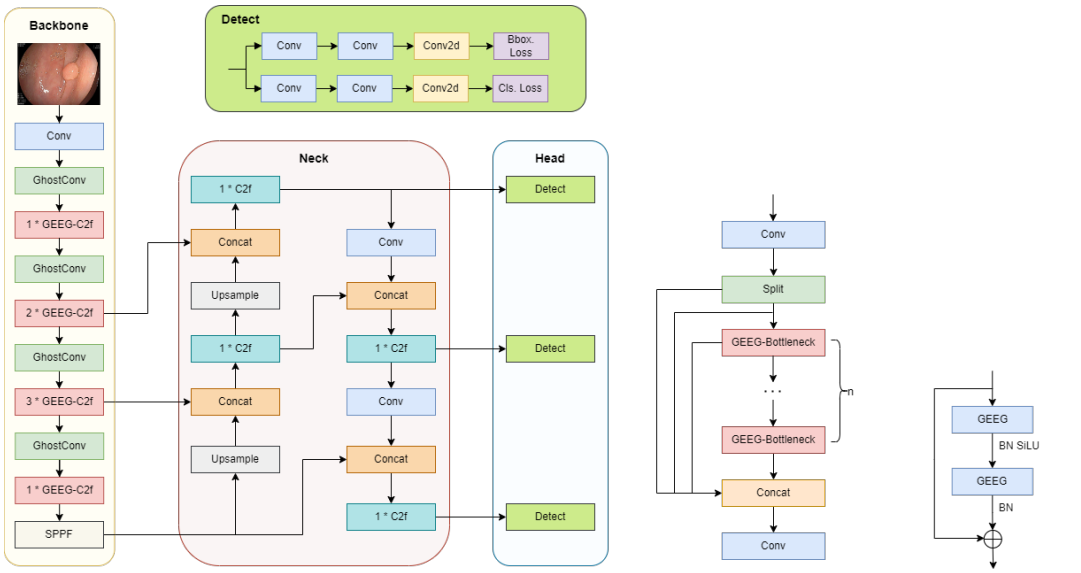
(a) Original image

(b) Gaussian Enhanced Euclidean norm Ghost attention module

(c) Ghost convolution

Fig. 1. 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.

The emergence of deep learning approaches has made a significant impact on the accuracy of automatic polyp detection on endoscopic images and videos. Nevertheless, there are factors that hinder the performance of such deep learning methods. The large differences in polyp characteristics such as shape, texture, and size pose difficulty for deep learning models in locating them. Polyps may be concealed by water flow, artifacts such as bubbles, light scatters during endoscopy procedure, and other bodily tissues. Under specific camera view, polyps appear very similar to intestine wall. Additionally, the endoscopy contains moving-camera, different from common moving object detection with fixed camera circumstances. The complex camera motion generates inevitable noises such as motion blur, occlusion, illuminance variations,



(a) GEEG-YOLOv8 (b) GEEG-C2f (c) GEEG-Bottleneck

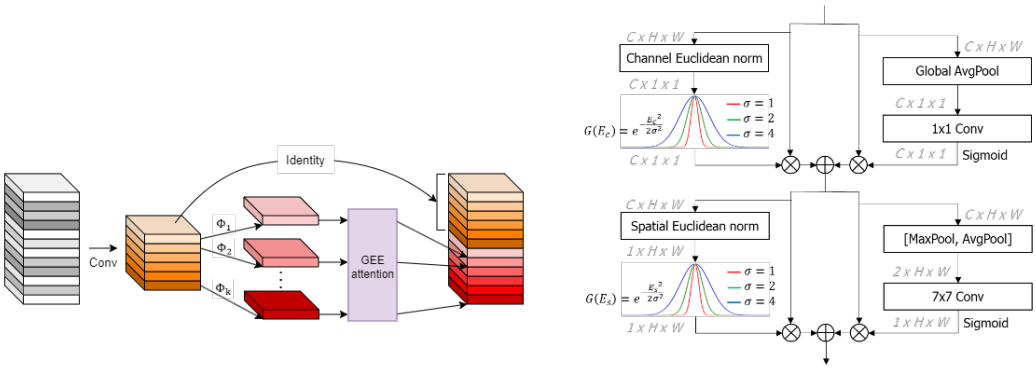
Fig. 2. The overall architecture of the proposed GEEG-YOLOv8.

etc., disturbing the detection results of deep learning models. This means that the model may miss the polyp in images or videos.

In recent years, many studies have been proposed to improve the performance of automatic polyp detection [2, 3, 4, 5, 6, 7, 8]. Most of them are video-based object detection which uses features from extra dimension. RYCO [2] and AIPDT [3] both incorporate temporal information by utilizing discriminate correlation filter-based trackers. Accurate detection of the polyp in the initial frame is required in these methods which is difficult to obtain due to noises in image. Zheng et al. [4] employs optical flow to refine detection result. However, high variation between consecutive video frames caused by complex movements of the camera leads to suboptimal performance. STFT [5] proposes Spatial-Temporal Feature Transformation to learn the alignment and mitigate the feature inconsistency between multiple frames. Although the detection accuracy of STFT increases, the heavy computational cost makes real-time running incapable. YONA [6] extract information from two consecutive frames for accurate and fast polyp detection by the presented foreground temporal alignment, background dynamic alignment, and cross-frame box-assisted contrastive learning module.

Regarding image-based object detection approaches, Shin et al. [7] applies region-based deep convolutional neural network and post-learning approaches to reduce the false positive rate of polyp detection. However, it also requires high computational resources, making real-time running infeasible. Wan et al. [8] introduces an attention mechanism to YOLOv5 model to achieve accurate and real-time polyp detection, but the model is evaluated on a private dataset and a small public dataset.

In this paper, a novel Gaussian Enhanced Euclidean norm Ghost attention (GEEG) module is presented for reliable real-time polyp detection on endoscopic images and videos. It does not require feature extraction from extra dimensions to tackle the above issues. By modifying the backbone of YOLOv8 model [9] to include Ghost convolution (GhostConv) [10], the number of model parameters and floats point operations (FLOPs) can be greatly reduced. Despite that, the use of GhostConv alone does not bring sufficient performance gain in polyp detection. The cheap operations in GhostConv, which is usually 3×3 depth-wise convolution, only capture the spatial information from inherent feature maps generated by 1×1 point-wise convolution, neglecting global dependency. The depth-wise convolution also does not consider the correlation between channel information. Therefore, cheap operations repeatedly extract local information produced from inherent feature maps, hindering performance improvement. To enhance the ability to extract inter-channel and spatial information of GhostConv, Gaussian Enhanced Euclidean norm (GEE) attention mechanism is added after cheap operations. This attention method is inspired by Convolutional Block Attention Module (CBAM) [11] and Gaussian Context Transformer (GCT) [12], and is based on a hypothesis of the relationship between global contexts information and attention activations: Smaller attention activations are linked to global contexts possessing larger absolute values [12]. 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, we use a Gaussian function which takes the Euclidean norm of



(a) Gaussian Enhanced Euclidean norm Ghost attention module (b) Gaussian Enhanced Euclidean norm attention

Fig. 3. Gaussian Enhanced Euclidean norm Ghost attention module block details. $\otimes$ represents broadcast element-wise multiplication and $\oplus$ denotes element-wise addition. $\mathbf{E_c}$ and $\mathbf{E_s}$ represents Channel Euclidean norm and Spatial Euclidean norm, respectively.

channel or spatial dimension as input to refine features output by the channel and spatial attention mechanism. As shown in Fig. 1, the feature maps generated by cheap operations in GEEG emphasize more important information, i.e. the polyp features, denoted by red arrow, compared to the naive GhostConv. GEEG's feature maps contain more fine-grained features, whereas one feature map generated by GhostConv captures noise information, i.e. light scatter, distracting the model.

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. We call it GEEG-YOLOv8, its architecture is shown in Fig. 2.
- Extensive experiments on three public datasets demonstrate superior performance compared to previous state-of-the-art methods.

2. PROPOSED METHOD

In this section, we will present the lightweight GEEG-YOLOv8 for high polyp detection performance. Then we will introduce the new GEEG module which mitigates the weakness of GhostConv in capturing global channel and spatial information.

2.1. GEEG-YOLOv8

The overall framework of GEEG-YOLOv8 is shown in Fig. 2a. Its architecture comprises a backbone, a neck, and a detection head. The backbone network is composed of multiple regular GhostConv and GEEG-C2f modules. Although GEEG can enhance the information extraction capability, it increases model complexity; therefore, GEEG is only incorporated in C2f to avoid introducing large number of parameters to the model. The GEEG-Bottleneck, as illustrated in Fig. 2c, contains two GEEG modules. The first GEEG module serves as a squeezing layer, reducing the number of input channels with a ratio of 2. The second GEEG module increases the number of channels to align with the shortcut path. Then the inputs and outputs of these two GEEG modules are linked by a residual connection. The first GEEG module is followed by Batch Normalization (BN) operation and SiLU activation function while only BN is used after the second GEEG module. The GEEG-C2f shown in Fig. 2b, includes n GEEG-Bottleneck which has two parallel gradient flow branches, thereby it can obtain richer gradient flow information while reducing parameters. The neck is PANet, which fuses feature maps from lower layers to deeper layers to boost information flow. GEEG-YOLOv8 adopts an anchor-free approach with a decouple detection head. The first branch in the detection head outputs object bounding box loss while the second branch outputs object class loss.

2.2. Gaussian Enhanced Euclidean norm Ghost attention module

2.2.1. Gaussian Enhanced Euclidean norm attention

Self-attention has been proven to be efficient in extracting long-range global dependencies. However, the quadratic complexity makes it impractical to implement in real time on constrained hardware devices. We propose the lightweight GEE attention to boost GhostConv performance with negligible number of additional parameters, based on a hypothesis: Smaller attention activations are attached to global contexts that have larger absolute values. GEE is composed of chan-

nel attention and spatial attention module, each module has two branches, as shown in Fig. 3b. The left branch differs from GCT in two ways. First, we directly take the Euclidean norm of the feature maps as input to a Gaussian function rather than using the normalization operation of global average pooling (GAP). The intuition behind is that the Euclidean norm quantifies the magnitude of a vector or matrix, with larger Euclidean norm indicating greater deviation from the vector or matrix to its origin, so constraining the Euclidean norm with Gaussian function will improve model generalizability. Second, we argue that the above hypothesis is also true in spatial dimension, hence we extend the idea to extract global spatial information. The right branch is quite similar to CBAM, except that in channel attention module, only GAP is used, followed by a 1×1 convolution layer to reduce model parameters.

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

$$\begin{aligned} \mathbf{F}' &= \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F} \oplus \mathbf{G}(\mathbf{E_c}) \otimes \mathbf{F} \\ \mathbf{F}'' &= \mathbf{M_s}(\mathbf{F}') \otimes \mathbf{F}' \oplus \mathbf{G}(\mathbf{E_s}) \otimes \mathbf{F}', \end{aligned} \tag{1}$$

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

$$\mathbf{E_c} = \left\{ e_{ck} = \sqrt{\sum_{i=1}^{W} \sum_{j=1}^{H} \mathbf{F}_k(i,j)^2} : k \in \{1, \ldots, C\} \right\} \tag{2}$$

$$\mathbf{E_s} = \left\{ e_{sij} = \sqrt{\sum_{k=1}^{C} \mathbf{F}'_{ij}(k)^2} : i \in \{1, \ldots, W\}, j \in \{1, \ldots, H\} \right\}. \tag{3}$$

A Gaussian function $\mathbf{G}(x) = \exp(-\frac{x^2}{2\sigma^2})$ takes x as input with maximum value 1, mean 0 and standard deviation σ, to satisfy the hypothesis about the relationship between global contexts and attention activations. Larger standard deviation results in smaller difference between each attention activation. The impact of standard deviation on model performance will be analyzed in the experiment part.

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

$$\mathbf{M_c}(\mathbf{F}) = \text{Sigmoid}(f^{1\times1}(\text{GAP}(\mathbf{F}))), \tag{4}$$

$$\mathbf{M_s}(\mathbf{F}) = \text{Sigmoid}(f^{7\times7}([\text{MaxPool}(\mathbf{F}); \text{AvgPool}(\mathbf{F})])), \tag{5}$$

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

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.

Table 1. Detection result of GEEG-YOLOv8 on SUN dataset with different standard deviation σ.

σ	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

2.2.2. *Combine with Ghost Convolution*

GEE attention is added after cheap operations in GhostConv to create GEEG module, as shown in Fig. 3a. The conventional GhostConv can reduce computation cost, but it has limited information extraction ability due to cheap operations only capturing local information from inherent feature maps generated by 1×1 point-wise convolution. GEE attention can elevate the capacity of GhostConv for extracting long-range global channel and spatial information. Given an input feature map $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$, GEEG performs two steps. First, the inherent feature map $\mathbf{Y}' \in \mathbb{R}^{C' \times H \times W}$ is generated by

$$\mathbf{Y}' = f^{1\times1}(\mathbf{F}). \tag{6}$$

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

$$\mathbf{Y} = \text{Concat}([\mathbf{Y}', \text{GEE}(\Phi_{dp}(\mathbf{Y}'))]), \tag{7}$$

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

3. EXPERIMENTS

3.1. Datasets

Following datasets are used to conduct our experiments: Kvasir-SEG [17], NeoPolyp-Small [18], PolypsSet [19], LDPolypVideo [20], CVC-ClinicDB [21], ETIS-LaribPolypDB [21], and SUN [22]. Considering the generalization capability of our proposed method, we combine the first four datasets as training set and test our model on the last three datasets. We manually delete images that have identical viewpoint and distance of the same polyp and images that are too blurry and contain too many artifacts in the training set since these images do not contribute meaningful information to the learning process. Totally, there are 24,734 training images. About the test datasets, CVC-ClinicDB and ETIS-LaribPolypDB has 612 and 196 polyp images, respectively. SUN database contains 49,136 polyp frames taken from different 100 polyp video sequences.

3.2. Experimental Setup

All experiments were conducted on one NVIDIA RTX 4070 GPU with 12GB VRAM. The models are deployed in PyTorch framework. SGD optimizer is utilized with an initial

Table 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 [5]	81.50	71.45	80.69	40.12	-	-	-	-	-	-	-	-
AIDPT [3]	80.37	69.78	79.21	38.55	-	-	-	-	-	-	-	-
YONA [6]	83.30	71.52	81.43	41.89	-	-	-	-	-	-	-	-
Wan et al [8]	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 [13]	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 [14]	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 [15]	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 [16]	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 [9]	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 3. Model complexity and frames per second rate comparisons.

Method	Params	GFLOPs	FPS
STFT [5]	-	-	12.5
AIDPT [3]	-	-	72
YONA [6]	-	-	46.3
Wan et al. [8]	-	-	45
EfficientDet-D0 [13]	3.9M	2.4	209
YOLOv3-tiny [14]	8.6M	12.9	370
YOLOv6s [15]	18.5M	45.2	275
YOLOv7-tiny [16]	6.0M	13.0	208
YOLOv8s [9]	11.1M	28.4	300
GEEG-YOLOv8 (ours)	8.5M	21.2	303

learning rate of 0.01 and momentum of 0.937, weight decay rate is 0.0005. All images are resized to 640×640, batch size is set to 32. The models were trained for 400 epochs. The detection performance is evaluated by precision (P), recall (R), mean average precision (IoU = 0.5) (mAP@50), and mean average precision (IoU ranging from 0.5 to 0.95) (mAP@50-95) metric. Models complexity and speed are measured by the number of parameters, GFLOPs, and frames per second (FPS) rate.

3.3. Experimental Results

3.3.1. Impact of Standard Deviation on Model Performance

In this section, the effect of standard deviation σ in Gaussian function $\mathbf{G}(x) = \exp(-\frac{x^2}{2\sigma^2})$ on GEEG-YOLOv8's detection performance is examined. The results are illustrated in Table 1. We can see that as σ increases, the network performance initially improves before subsequently declining. The best score is obtained when σ is set to 4. This pattern is logical because too large variance diminishes the difference in attention activations among channel and spatial dimension, hindering the effective suppression of global noise contexts, while too small variance may overly restrain the importance of other meaningful features as well as wrongly highlight noise contexts.

3.3.2. Comparisons with State-of-the-arts Methods

Table 2 shows the quantitative comparison results of the proposed GEEG-YOLOv8 with other detection models. Overall, our model outperforms other methods in all metrics on SUN and CVC-ClinicDB dataset. On ETIS-LabribPolypDB dataset, GEEG-YOLOv8 obtains the best mAP@50 and mAP@50-95, and second best P and R. Compared with runner-up models, GEEG-YOLOv8 achieves great performance gains of 3% in P and mAP@50-95 metric on SUN, 0.5% in R and mAP@50 score on CVC-ClinicDB, and 2% on mAP@50 and mAP@50-95 metric on ETIS-LaribPolypDB. The proposed GEEG-YOLOv8 outperforms other models specialized in polyp detection. Model complexity and FPS rate are illustrated in Table 3. With the efficient and lightweight design of GEEG, the number of parameters and GFLOPs of our model are reduced compared to the original YOLOv8 model. Despite not having the lowest number of parameters and GFLOPs, GEEG-YOLOv8 attains the second-best FPS of 303. Our proposed method achieves new state-of-the-art polyp detection accuracy with low model complexity, showing its potential in real-life applications where real-time running is crucial.

Fig. 4 shows the qualitative results of GEEG-YOLOv8 with other models on three datasets. Our proposed method can correctly detect polyps under various conditions, such as blurring effect in SUN dataset. It is also capable of identifying flat and small polyps in CVC-ClinicDB and ETIS-LaribPolypDB. Moreover, GEEG-YOLOv8 shows a robust performance for polyps that look similar to intestine wall.

3.3.3. Ablation Study

To investigate the contribution of different components in the proposed method, ablation experiments were conducted on SUN dataset. The results are shown in Table 4. In method (a), only GhostConv is used in the backbone of the original YOLOv8. Although method (a) has the highest FPS, the detection performance is sub-optimal due to the limitation in extracting global information ability of GhostConv. Method (b) adds the channel and spatial attention mechanism (the right

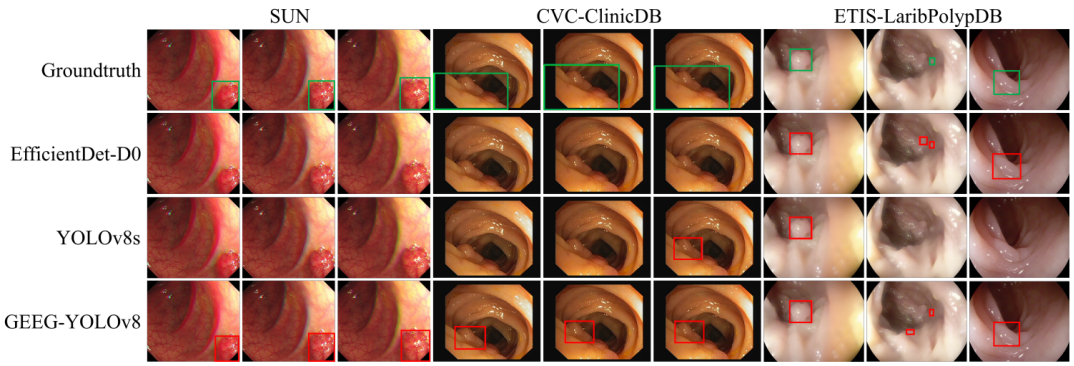


Fig. 4. Qualitative comparisons of polyp detection on three datasets. Green and red denotes the groundtruth and model prediction, respectively.

Table 4. Ablation results on SUN dataset of the proposed method.

Method	GhostConv	$\mathbf{M}(x)$	$\mathbf{G}(x)$	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

branch in Fig. 3a) after cheap operations in GhostConv, obtaining 3% gains in R, mAP@50, and mAP@50-95 metric with only 0.1M more parameters. Method (c) applies only GEE (the left branch in Fig. 3a) to GhostConv, and compared to method (b), about 1% increase in all four detection metrics is attained without additional parameters. By incorporating all three improvements in method (d), the best polyp detection score is achieved with no additional parameters and GFLOPs compared to method (b), and only a small reduction in FPS rate is shown. The P, R, mAP@50, and mAP@50-95 score of method (c) is 87.23%, 71.56%, 82.49%, and 45.41%, respectively.

4. CONCLUSION

In this paper, we propose a novel Gaussian Enhanced Euclidean norm Ghost attention module to improve the performance of polyp detection on endoscopic images and videos in both accuracy and speed. It applies a Gaussian function on the Euclidean norm of channel and spatial dimension to refine features output by the channel and spatial attention mechanism, improving the ability of Ghost convolution in extracting global long-range context information. This module is integrated into the backbone of YOLOv8 to form a new model named GEEG-YOLOv8. Extensive experimental results demonstrate that our proposed method shows a strong generalization capability without the need to extract information from extra dimension. With a small gain in number of model parameters and GFLOPs, GEEG-YOLOv8 achieves state-of-the-art polyp detection performance on three datasets. In the future, we will work on modifying the neck and detection head to further improve the detection result of our model.

5. REFERENCES

[1] Johannes Asplund, Joonas H Kauppila, Fredrik Mattsson, and Jesper Lagergren, "Survival trends in gastric adenocarcinoma: a population-based study in sweden," *Annals of surgical oncology*, vol. 25, pp. 2693–2702, 2018.

[2] Ruikai Zhang, Yali Zheng, Carmen CY Poon, Dinggang Shen, and James YW Lau, "Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker," *Pattern recognition*, vol. 83, pp. 209–219, 2018.

[3] Zijian Zhang, Hong Shang, Han Zheng, Xiaoning Wang, Jiajun Wang, Zhongqian Sun, Junzhou Huang, and Jianhua Yao, "Asynchronous in parallel detection and tracking (aipdt): Real-time robust polyp detection," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*. Springer, 2020, pp. 722–731.

[4] 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 Sympo-*

sium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 79–82.

[5] Lingyun Wu, Zhiqiang Hu, Yuanfeng Ji, Ping Luo, and Shaoting 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: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 302–312.

[6] Yuncheng Jiang, Zixun Zhang, Ruimao Zhang, Guanbin Li, Shuguang Cui, and Zhen Li, "Yona: You only need one adjacent reference-frame for accurate and fast video polyp detection," *arXiv preprint arXiv:2306.03686*, 2023.

[7] Younghak Shin, Hemin Ali Qadir, Lars Aabakken, Jacob Bergsland, and Ilangko Balasingham, "Automatic colon polyp detection using region based deep cnn and post learning approaches," *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[8] Jingjing Wan, Bolun Chen, and Yongtao Yu, "Polyp detection from colorectum images by using attentive yolov5," *Diagnostics*, vol. 11, no. 12, pp. 2264, 2021.

[9] Glenn Jocher, Ayush Chaurasia, and Jing Qiu, "Ultralytics YOLO," Jan. 2023.

[10] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1580–1589.

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[12] Dongsheng Ruan, Daiyin Wang, Yuan Zheng, Nenggan Zheng, and Min Zheng, "Gaussian context transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15129–15138.

[13] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781–10790.

[14] Joseph Redmon and Ali Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[15] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al., "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[16] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.

[17] 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 *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[18] Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang, "Neounet: Towards accurate colon polyp segmentation and neoplasm detection," in *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*. Springer, 2021, pp. 15–28.

[19] Kaidong Li, Mohammad I Fathan, Krushi Patel, Tianxiao Zhang, Cuncong Zhong, Ajay Bansal, Amit Rastogi, Jean S Wang, and Guanghui Wang, "Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations," *Plos one*, vol. 16, no. 8, pp. e0255809, 2021.

[20] Yiting Ma, Xuejin Chen, Kai Cheng, Yang Li, and Bin Sun, "Ldpolypvideo benchmark: a large-scale colonoscopy video dataset of diverse polyps," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, 2021, pp. 387–396.

[21] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[22] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Kinichi Hotta, Kazuo Ohtsuka, Takahisa Matsuda, Shoichi Saito, Toyoki Kudo, Toshiyuki Baba, Fumio Ishida, et al., "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal endoscopy*, vol. 93, no. 4, pp. 960–967, 2021.