Feature Selection Gates with Gradient Routing for Endoscopic Image Computing

Giorgio Roffo [1]✉, Carlo Biffi [1]
Pietro Salvagnini [1], and Andrea Cherubini [1,2]✉

[1] Cosmo Intelligent Medical Devices, Dublin, Ireland
[2] Milan Center for Neuroscience, University of Milano–Bicocca, Milano, Italy
{groffo, acherubini}@cosmoimd.com

Abstract. To address overfitting and enhance model generalization in gastroenterological polyp size assessment, our study introduces Feature Selection Gates (FSG) alongside Gradient Routing (GR) for dynamic feature selection. This technique aims to boost Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) by promoting sparse connectivity, thereby reducing overfitting and enhancing generalization. FSG achieves this through sparsification with learnable weights, serving as a regularization strategy. GR further refines this process by optimizing FSG parameters via dual forward passes, independently from the main model, to improve feature re-weighting. Our evaluation spanned multiple datasets, including CIFAR-100 for a broad impact assessment and specialized endoscopic datasets (REAL-Colon [13], Misawa [10], and SUN [14]) focusing on polyp size estimation, covering over 200 polyps in more than 370K frames. The findings indicate that our FSG-enhanced networks substantially enhance performance in both binary and triclass classification tasks related to polyp sizing. Specifically, CNNs experienced an F1 Score improvement to 87.8% in binary classification, while in triclass classification, the ViT-T model reached an F1 Score of 76.5%, outperforming traditional CNNs and ViT-T models. To facilitate further research, we are releasing our codebase, which includes implementations for CNNs, multistream CNNs, ViT, and FSG-augmented variants. This resource aims to standardize the use of endoscopic datasets, providing public training-validation-testing splits for reliable and comparable research in gastroenterological polyp size estimation. The codebase is available at github.com/cosmoimd/feature-selection-gates.

Keywords: Feature Selection Gates · Gradient Routing · Attention Gates · Endoscopy · Medical Image Processing · Computer Vision

1 Introduction

Deep learning (DL) techniques, particularly convolutional neural networks and transformers, have significantly advanced the analysis of endoscopic imaging [3] [18] [22] [25]. Nonetheless, these models encounter challenges such as overfitting when applied to the typically smaller datasets found in endoscopy, in contrast

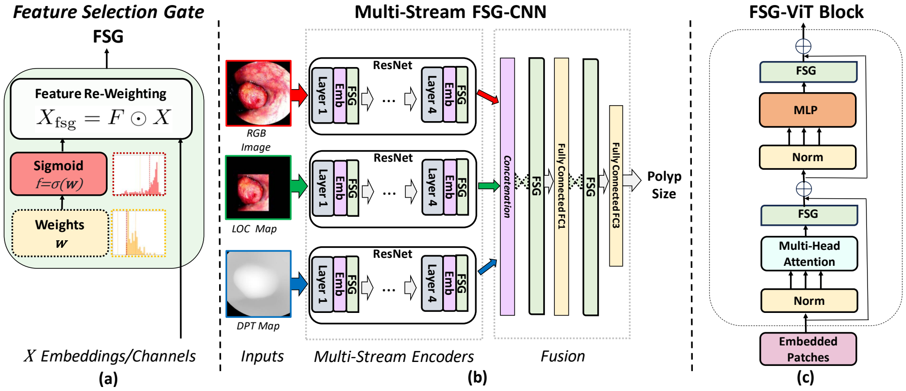


Fig. 1. Feature Selection-Attention Gates (FSG) Integration in Deep Learning Models. (a) Conceptual design of FSG. (b) Application of FSG in a multi-stream CNN architecture, with each stream being optional. (c) Embedding of FSG within a ViT block, positioned after the multihead attention and the MLP for enhanced feature reweighting.

to larger datasets like ImageNet. These challenges stem from various factors, including privacy concerns, the need for expert annotations, associated costs, and the inherent variability of endoscopy imaging modalities. Moreover, the infrequency of certain conditions, such as large colorectal polyps, intensifies data imbalance issues, further complicating the development of reliable and precise DL models for endoscopic image analysis [7] [8] [15] [2] [4] [23].

A review of related work on polyp size estimation underscores the critical nature of accurate assessment for the effective management and surveillance of colorectal cancer [35]. The size of a polyp significantly influences its potential for malignancy, necessitating accurate measurement. According to guidelines from prominent endoscopy organizations, including the American and European Societies for Gastrointestinal Endoscopy (*ASGE* and *ESGE*), polyps can be classified in D-S-L categories based on their size: (D) Diminutive polyps (5mm or smaller), (S) Small polyps (between 5mm and 10mm), and (L) Large polyps (10mm or larger). The societies particularly emphasize the removal and subsequent histopathological examination of Large polyps due to their increased risk of cancer, underscoring the critical role of precise size determination [32,33]. However, there is notable variability in manual size estimations by endoscopists [4], which poses a risk of mismanagement [34]. The development of automated estimation techniques via computer vision aims to mitigate this issue, providing more consistent measurements through machine learning models trained on a diverse array of imaging data [35]. In pursuit of standardizing polyp size estimation, our work leverages state-of-the-art 2D and 3D estimation methodologies [2,15,6,5] and enhances dataset diversity [13,10,14]. This dual approach aims to refine the accuracy of automated polyp-size estimation, addressing both the technological and data-related challenges inherent in this domain.

Building upon this foundation, we introduce Feature Selection-Attention Gates (FSG) and Gradient Routing (GR). These innovative mechanisms, tailored for the gastroenterological polyp size assessment domain, counter overfitting and enhance model generalization. Drawing from [19,17,26,21,20], our approach promotes sparse connectivity in deep networks and uses a dual forward pass strategy for gradient routing. This fosters model sparsity and efficiency while selectively emphasizing pertinent features [19,1]. Figure 1 illustrates the integration of FSG in DL frameworks. Fig. 1(a) outlines the FSG's conceptual design, employing sigmoid-normalized weights ranging between 0 and 1. Fig. 1(b) integrates FSG into CNNs in a multi-stream setup, accommodating various input types like RGB, Depth, and Location Maps. Fig. 1(c) illustrates FSG's incorporation in the ViT model, placing one FSG after the multihead attention and another following the MLP block.

The proposed approach was evaluated, with our primary focus on public endoscopic datasets, consisting of 232 polyps across more than 370K frames in the REAL-Colon [13], Misawa [10] and the SUN database [14]. These databases encompass realistic clinical scenarios, such as variable lighting and obstructions. Additionally, to assess the impact of FSG on ViT performance in a more general context, we also conducted evaluations on CIFAR-100. The FSG ResNet18 (R18) and ViT models showed improved accuracy, achieving an accuracy of 75.2% and 83.8% respectively, outperforming by +1.3% and +5.8% their baselines and SotA [30,28]. In polyp size estimation experiments, we compared methods with SotA using RGB, Depth [31], and Location Maps in CNNs with Dropout and Batch Normalization. Hybrid methods like CBAM [36] were excluded to focus on pure CNN and transformer architectures.

In the bias-variance tradeoff analysis, FSG models showed superior performance with higher F1 scores and average sensitivity-specificity in a 6-fold experiment. In a separate, consolidated evaluation across all dataset folds, FSG models consistently outperformed standard models in both binary [7,8] (under and over 10 mm polyps) and triclass classifications (Diminutive-Small-Large).

Across 12 methods compared, the highest average performances were from **FSG MultiStream-R18 (LOC+DPT)** at **66.1%** and **FSG ViT-T (RGB)** at **65.5%**. Our findings indicate that ViT models without LOC maps or DPT are preferred due to reduced error propagation. ViT Tiny, with 5.6M parameters and 4.7G flops, is the most efficient. FSG integration significantly enhances ViT's performance in regression and classification. Our unique integration of the following components to address overfitting in medical image analysis is a key innovation. The main contributions of our work can be summarized as follows:

1. **Feature Selection Gate (FSG)**: Acts as an online regularization tool to enhance learning, reduce overfitting, and improve generalization.
2. **Gradient Routing (GR)**: Optimizes FSG parameters separately from the main model, allowing tailored learning rates and gradient clippings.
3. **Enhancing Vision Transformers and CNNs**: FSG enhances ViTs and CNNs, including multi-input variants, for versatile processing of RGB, Depth maps [31], and location maps.

2 Methodology

2.1 Feature Selection-Attention Gates (FSG)

The FSG (in Fig.1) dynamically assigns weights to each feature within the model, applicable to embeddings and channels in architectures such as Transformers and CNNs. These weights, dynamically adjustable during the training process, are normalized using a sigmoid function to ensure values range between 0 and 1. This weighting mechanism facilitates focused learning, allowing the model to prioritize more informative features and reduce the less relevant ones.

Given a set of input features X, represented as $X = (x_1, x_2, \ldots, x_n)$, where n is the number of features or the embedding size, and each x_i for $i = 1, 2, \ldots, n$ corresponds to a specific feature in the input data. Feature selection is performed by introducing a set of weights F, denoted as $F = (f_1, f_2, \ldots, f_n)$. These weights are derived from the raw feature weights $W = (w_1, w_2, \ldots, w_n)$ through a transformation. Specifically, we apply a sigmoid function σ to each raw weight w_i to obtain the corresponding FS weight f_i. The sigmoid function, defined as $\sigma(z) = \frac{1}{1+e^{-z}}$ for any input z makes the weights suitable for FS by scaling them between 0 and 1. These transformed weights are referred to as FSG-scores.

The FSG-scores are then applied to the original input features X to obtain the relevant version of the input, denoted as X_{fsg}. This application is performed through the Hadamard product of F and X, effectively scaling each feature x_i by its corresponding FSG-score f_i (see Fig.1). Mathematically, this process is encapsulated in the following equation:

$$\begin{aligned} X_{\text{fsg}} &= F \odot X = f_1 \cdot x_1, f_2 \cdot x_2, \ldots, f_n \cdot x_n \\ &= \sigma(w_1) \cdot x_1, \sigma(w_2) \cdot x_2, \ldots, \sigma(w_n) \cdot x_n, \end{aligned} \tag{1}$$

where $\odot$ denotes the Hadamard product, and each $f_i = \sigma(w_i)$ is the result of applying the sigmoid function to the raw weight w_i, which is then multiplied by the corresponding input feature x_i to achieve the feature selection effect.

Unlike attention mechanisms that use softmax for input weighting, FSG *independently* re-weights features with scores from 0 to 1, not summing to 1. This enables synergy with attention in ViT architectures, enhancing performance (see FSG-GR weight distributions in supplementary material).

2.2 Gradient Routing for Online Feature Selection

Gradient Routing (GR) in our model employs a dual-phase optimization approach with distinct optimizers for different model components. Initially, GR updates the FSG parameters. This step focuses on refining feature weights only. Following this, GR updates the main model parameters using a different optimizer, based on the adjusted state from the FSG phase. The iterative nature of GR aligns with the principles of gradient descent and backpropagation, starting with the fine-tuning of FSG parameters and then progressing to the main model parameters. The small derivatives introduced by the sigmoid function in deep layers can lead to vanishing gradients, minimally updating early layer weights. To counter this, a gradient clipping strategy with different thresholds for FSG and the main model can be used. Higher thresholds for FSG address the sigmoid's limitations, and lower ones for the main model ensure stability. This method ensures efficient backpropagation across the network [24], optimizing learning in both FSG and the main model components. The GR method utilizes a dual-phase optimization with gradient clipping for FSG and main model parameters, diverging from the layer-wise pre-training and fine-tuning strategy described in [27]. The gradient updating process in GR can be represented as:

$$\theta_{fsg}^{t+1} = \theta_{fsg}^{t} - \eta_{\text{fsg}}\text{clip}(\nabla_{fsg} L(\theta_{fsg}^{t}, \theta_{main}^{t}, D), \text{Th}_{fsg}) \tag{2}$$

$$\theta_{main}^{t+1} = \theta_{main}^{t} - \eta_{main}\text{clip}(\nabla_{main} L(\theta_{main}^{t}, \theta_{fsg}^{t+1}, D), \text{Th}_{main}) \tag{3}$$

where $\theta_{\text{main}}^{(t)}$ and $\theta_{\text{fsg}}^{(t)}$ are the parameters of the main model and FSG at iteration t, η_{main} and η_{fsg} are the respective learning rates, ∇L denotes the gradient of the loss function, D is the training data, and $\text{clip}(\cdot, \text{Th})$ is the gradient clipping function with specified thresholds.

3 Experiments and Results

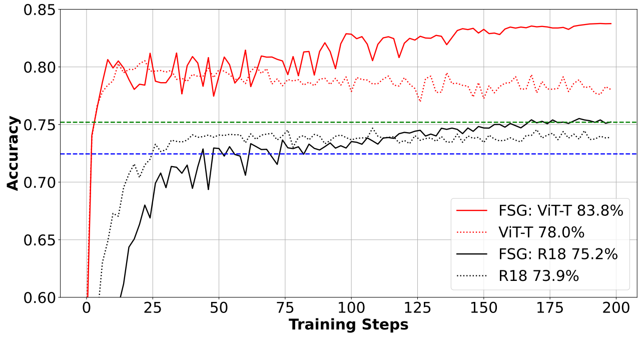
Images were resized to 384 × 384 and normalized using dataset-specific mean and standard deviation computed on the training data, ensuring dataset-specific color adjustments. Circular cropping was used to isolate the central part of an image into a circular shape, thereby concentrating analysis on relevant areas and eliminating peripheral distractions. Standard data augmentation included rotations, color adjustments, and noise addition. Polyp sizes normalized to $[-1, +1]$ range were used for stable regression training. A domain-specific weighted Huber Loss addressed the imbalanced distribution of polyp sizes within the dataset:

$$\mathbf{A} = \begin{cases} \alpha_1, & \text{if } T_1 < y \leq T_2 \\ \alpha_2, & \text{if } y > T_2 \\ 1, & \text{otherwise} \end{cases} \qquad L_W = \tfrac{1}{N}\textstyle\sum_{i=1}^{N}(\text{Huber}(x_i, y_i) \cdot \mathbf{A}_i) \tag{4}$$

with $(T_1, \alpha_1) = (5, 2)$ and $(T_2, \alpha_2) = (10, 3)$, and N representing the mini-batch size. The Adam optimizer was utilized, with a learning rate set in the range of $[10^{-3}, 10^{-5}]$, and weight decay specified within the interval $[10^{-5}, 10^{-8}]$. Gradient clipping was set between 5 to 10 for CNNs, and a cosine annealing scheduler with warm restarts was applied for learning rate control. Parameters within the FSG modules were initialized using the Xavier method. For ViT models, gradient clipping threshold was set to 128 to mitigate the vanishing gradient issue [24] (see Supplementary Material for experimental setup). We reported model parameters in Table 2 and conducted experiments using a Tesla V100-PCIE GPU with 32GB memory. ResNet-18 and ViT-Tiny models operate in real-time, requiring approximately 2GB for training and inference with a batch size of 1

Table 1. Distribution of Frames (#Polyps) in Endoscopic Dataset Folds from REAL-Colon [13], Misawa et al.'s Database [10], and the SUN Dataset [14]

Categories	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Fold 6**
Diminutive	46,982 (29)	49,648 (30)	44,640 (30)	54,707 (30)	49,256 (29)	51,784 (29)
Small	9,535 (5)	11,985 (6)	6,098 (6)	827 (5)	1,255 (6)	4,538 (4)
Large	6,298 (5)	24,189 (3)	1,086 (3)	6,459 (4)	1,105 (3)	1,648 (5)
Total	62,815 (39)	85,822 (39)	51,824 (39)	61,993 (39)	51,616 (38)	57,970 (38)



Methods	Acc.
R18 in [28]	74.1%
R18 (re-train)	73.9%
R18+FSG	**75.2%**
ViT-T in [30]	75.2%
ViT-T (re-train)	78.0%
ViT-T+FSG	**83.8%**

Fig. 2. ViT/R18 Performance on CIFAR-100 compared with SotA [30,28].

on 384x384x3 images. The FSG adds minimal parameters and an imperceptible increase in FLOPS, as detailed in Table 2. The experiments were conducted using the CIFAR-100, providing a standard benchmark for assessing classification accuracy (100 classes, 50K training images, 10K test images) and three endoscopic databases, namely, the REAL-Colon [13], Misawa et al.'s database [10], and the SUN dataset [14], consisting of a total of 232 unique polyps, represented by 372,040 frames. This dataset is partitioned into six folds to facilitate k-fold cross-validation. The database's overview, detailed in Table 1, highlights the distribution of lesions in the dataset, revealing a significant imbalance among the various types of lesions; for more specifics, please see [13,10,14]. For the sizing task, we used a 6-Fold cross-validation, allocating one fold each for testing and validation, and four folds for training in each cycle.

Exp. 1: Evaluating the ViT-Tiny and ResNet-18 on CIFAR-100

Our experimental analysis on the CIFAR-100 dataset highlights the significant impact of integrating FSGs with CNN and Transformer architectures, specifically ResNet-18 (R18) and Vision Transformer Tiny (ViT-T), on classification accuracy. Initially, the R18 model reached a 73.9% accuracy, aligning with previous benchmarks [28]. Incorporation of FSG into R18 improved its accuracy to 75.2%, indicating a 1.8% improvement. For ViT-T, the initial accuracy stood at 78.0%, comparable to standards set in [30]. However, applying FSG to ViT-T significantly increased its accuracy to 83.8%, marking a substantial 5.8% en-

Method Name	**Balanced Accuracy**		**Avg. Sens-Spec**		**F1 Score**		**Global Average**	**Params**	**#flops**
	Value	**Variance**	**Value**	**Variance**	**Value**	**Variance**	**of Metrics**		
R18 [16,28](RGB)	**53.4**%	1.4%	57.3%	1.0%	77.5%	**1.8**%	62.73%	11.2M	5.3G
FSG: R18 (RGB)	52.3%	**0.7**%	**57.9**%	**0.5**%	**78.7**%	2.0%	**62.97**%	11.3M	5.3G
R18 [12,31](DPT)	**46.9**%	1.0%	46.8%	1.6%	66.2%	5.0%	53.30%	11.2M	5.1G
FSG: R18 (DPT)	46.8%	**0.4**%	**50.1**%	**0.3**%	**76.5**%	**2.1**%	**57.80**%	11.2M	5.1G
R18 (LOC)	**54.3**%	1.7%	**56.4**%	1.3%	**78.1**%	2.3%	**62.93**%	11.2M	5.3G
FSG: R18 (LOC)	53.0%	**1.2**%	55.4%	**0.9**%	77.6%	**1.7**%	62.00%	11.3M	5.3G
MultiStream-R18 [9][RGB+DPT]	50.7%	1.3%	54.9%	**0.8**%	77.0%	**2.2**%	60.87%	22.4M	5.3G
FSG: MultiStream-R18 (RGB+DPT)	**53.4**%	1.3%	**57.1**%	0.9%	**77.8**%	2.5%	**62.77**%	22.5M	5.3G
MultiStream-R18 [9][LOC+DPT]	52.0%	**0.7**%	53.1%	**0.4**%	78.2%	1.7%	61.10%	22.4M	5.3G
FSG: MultiStream-R18 (LOC+DPT)	**53.5**%	0.9%	**56.0**%	0.5%	**79.5**%	**1.4**%	**63.00**%	22.5M	5.3G
ViT-Tiny [11,30](RGB)	51.3%	1.2%	55.7%	**0.6**%	75.6%	2.0%	60.87%	**5.6M**	**4.7G**
FSG: ViT-Tiny (RGB)	**54.9**%	**1.1**%	**59.5**%	0.7%	**79.1**%	2.0%	**64.50**%	**5.6M**	**4.7G**

Table 2. Bias & Variance: Diminutive (<=5mm), Small (5-10mm), Large (>=10mm)

hancement over the baseline. These improvements underscore FSG's capability to selectively emphasize influential features, thereby optimizing model performance across both architectures. This foundational assessment sets the stage for applying FSG in more specialized tasks like polyp size estimation, demonstrating its potential to refine accuracy in complex image classification challenges.

Exp. 2: Polyp-Size Estimation. Bias-Variance Tradeoff Analysis

In our experimental setup, models were trained on four folds, with one fold each for validation and testing, as per Table 1. The optimal model checkpoint is chosen based on the lowest validation loss.

Table 2 presents a detailed bias-variance tradeoff analysis for models enhanced with FSGs using RGB, LOC, and DPT inputs. The integration of FSG into the ViT model with RGB inputs increases Balanced Accuracy (BA) from 51.3% to 54.9%, improves Avg. Sensitivity-Specificity to 59.5%, and raises the F1 Score to 79.1%. This results in a global average metric improvement of +3.63% relative to its non-FSG counterpart, achieving 64.5%. In the case of the R18 model equipped with DPT inputs, the integration of FSG results in an increase of Avg. Sensitivity-Specificity by +3.3% and an F1 Score by +10.3%, both compared to the model's performance without FSG. Applying FSG to the R18 model with LOC inputs results in a slight decrease in BA by approximately 1%. The combination of RGB and polyp location masks, which utilize ground truth (GT) bounding boxes to create the location masks, represents an ideal input with optimal feature selection by design. In this scenario, FSG had limited scope for reweighting and selection as most relevant information was already incorporated. For multi-stream configurations combining LOC and DPT inputs, the enhancements include a BA increase to 53.5% and the attainment of the highest F1 Score at 79.5%. These improvements are due to FSG and GR. FSG promotes sparse connectivity, reducing overfitting and improving generalization. GR optimizes FSG with dual forward passes, focusing on key features and eliminating redundancies when the main model parameters are frozen. This ensures the model focuses on relevant features, enhancing predictive accuracy and robustness across input modalities.

Method Name	Binary Classification				Triclass Classification				Overall
	Bal. Acc.	**F1 Score**	**Sens.-Spec.**	**Avg.**	**Bal. Acc.**	**F1 Score**	**Sens.-Spec.**	**Avg.**	**Score**
R18 [16,28](RGB)	**57.58%**	**87.22%**	**70.20%**	71.67%	**47.62%**	74.84%	**54.18%**	58.88%	65.2%
FSG: R18 (RGB)	55.54%	86.46%	**70.10%**	70.70%	45.57%	**75.68%**	**54.00%**	58.42%	64.6%
R18 [12,31](DPT)	**58.20%**	78.93%	56.51%	64.55%	44.19%	68.12%	42.72%	51.68%	58.1%
FSG: R18 (DPT)	54.75%	**85.84%**	**63.30%**	67.96%	**45.0%**	**73.58%**	**49.00%**	55.86%	61.9%
R18 (LOC)	54.12%	85.59%	62.56%	67.42%	44.76%	74.99%	48.93%	56.23%	61.8%
FSG: R18 (LOC)	**55.70%**	**86.24%**	**64.80%**	68.91%	**45.42%**	74.99%	**50.11%**	56.84%	62.9%
MultiStream-R18 [9][RGB+DPT]	**54.40%**	**85.90%**	**67.09%**	69.13%	44.36%	73.81%	51.34%	56.50%	62.8%
FSG: MultiStream-R18 [RGB+DPT]	53.63%	85.55%	65.49%	68.22%	**45.81%**	**74.36%**	**52.01%**	57.39%	62.8%
MultiStream-R18 [9][LOC+DPT]	55.92%	86.16%	63.36%	68.48%	47.00%	75.68%	50.25%	57.64%	63.1%
***FSG: MultiStream-R18 [LOC+DPT]**	**59.96%**	**87.84%**	**69.23%**	72.34%	**48.84%**	**77.11%**	**53.56%**	**59.84%**	**66.1%***
ViT-Tiny [11,30](RGB)	54.63%	86.04%	68.35%	69.67%	42.78%	72.80%	50.82%	55.47%	62.5%
***FSG: ViT-Tiny (RGB)**	**55.86%**	**86.56%**	**69.33%**	70.58%	**48.93%**	**76.47%**	**55.62%**	**60.34%**	**65.5%***

Table 3. Performance Summary: *Binary* [7] (Polyps < 10mm vs. >= 10mm) **vs** *Triclass Classification* (Diminutive: <= 5mm, Small: 5-10mm, Large: >= 10mm)

Exp. 3: Comprehensive Model Performance Analysis

In our analysis, models enhanced with FSGs were evaluated for generalizability across *binary* and *triclass* classifications, as detailed in Table 3. In this experiment, we consolidated all inferences across different folds for each model to provide a comprehensive overview of their performance.

In binary classification tasks, the FSG MultiStream-R18 (LOC+DPT) model achieved a Balanced Accuracy (BA) of 59.96%, an F1 Score of 87.84%, and a Sensitivity-Specificity of 69.23%, with an overall average performance of 72.34%. This model showcases the efficacy of FSG in improving precision and predictive accuracy, setting a high benchmark in the binary classification domain.

In the context of triclass classification, the task's complexity significantly escalates. Nonetheless, the FSG-enhanced ViT-T (RGB) model showcases notable performance, achieving a Balanced Accuracy (BA) of 48.93%, an F1 Score of 76.47%, and a Sensitivity-Specificity of 55.62%, culminating in an average of 60.34%. These metrics not only underscore the model's robustness but also its adaptability to more complex classification scenarios, despite having only 5M parameters. This is considerably less — four times fewer than the multistream networks and half that of the R18.

The performance of the FSG-enhanced models, particularly MultiStream-R18 (LOC+DPT) in binary classification and ViT-T (RGB) in triclass classification, underscores the efficacy of FSG in model optimization across different classification tasks. Comparing 12 methods, the highest performances were from **FSG MultiStream-R18 (LOC+DPT)** at **66.1%** and **FSG ViT-T (RGB)** at **65.5%**. ViT models without LOC maps or DPT are preferred, due to their lower probability of error propagation from detection and depth estimation frameworks. Moreover, ViT-Tiny has four times fewer parameters and lower FLOPs (5.6M vs. 22.5M and 4.7G vs. 5.3G) compared to MultiStream-R18 (LOC+DPT). Therefore, the most promising solution is *FSG ViT-Tiny* as shown in this paper; Vision Transformer models improve significantly in both CIFAR-100 (classification in natural imaging) and polyp size estimation (regression in medical imaging).

4 Conclusions

This study advances deep learning for polyp size assessment by innovatively integrating Feature Selection-Attention Gates (FSG) with Gradient Routing (GR) across CNN and ViT architectures. For polyp sizing, the FSG-enhanced MultiStream-R18 (LOC+DPT) model excels in binary classification, achieving an F1 Score of 87.8% and an average performance of 72.3%. In triclass classification, the ViT-T model attains an F1 Score of 76.5% and an average of 60.3%, highlighting its efficiency and adaptability. Furthermore, the FSG-enhanced ViT achieves 83.8% accuracy in CIFAR-100, demonstrating its versatility for various imaging tasks. ViT models without LOC maps or DPT are preferred due to their lower probability of error propagation in detection and depth estimation frameworks. Moreover, ViT Tiny, with 5.6M params and 4.7G flops, has the lowest parameter count. Integrating FSG enhances ViT, achieving top performance in regression and classification. Advanced Vision Transformers like Swin, DeiT, and PVT show significant potential for future research. This analysis emphasizes the pivotal role of FSG-GR in improving polyp size estimation, suggesting beneficial effects on clinical outcomes. We aim to expand the application of these techniques to a broader range of medical imaging challenges, improving diagnostic accuracy with minimal computational overhead. To facilitate future research, the dataset splits and codebase for CNNs, multistream CNNs, ViT, and FSG-enhanced models is available at github.com/cosmoimd/feature-selection-gates.

Disclosure of Interests. All the authors are affiliated with Cosmo Intelligent Medical Devices, the developer of the GI Genius medical device.

References

1. G. Roffo, S. Melzi, U. Castellani, A. Vinciarelli, M. Cristani. Infinite feature selection: a graph-based feature filtering approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4396–4410, 2020.
2. M. Abdelrahim, H. Saiga, N. Maeda, E. Hossain, H. Ikeda, P. Bhandari. Automated sizing of colorectal polyps using computer vision. *Gut*, 71(1):7–9, 2022.
3. W. Jin, R. Daher, D. Stoyanov, F. Vasconcelos. A Self-supervised Approach for Detecting the Edges of Haustral Folds in Colonoscopy Video. In *MICCAI W. on Data Engineering in Medical Imaging*, pages 56–66, 2023.
4. C. Atalaia-Martins, P. Marcos, C. Leal, S. Barbeiro, A. Fernandes, A. Santos, L. Eliseu, C. Gonçalves, I. Cotrim, H. Vasconcelos. Variation between pathological measurement and endoscopically estimated size of colonic polyps. *GE-Portuguese Journal of Gastroenterology*, 26(3):163–168, 2019.
5. V. M. Batlle, J. M. M. Montiel, P. Fua, J. D. Tardós. LightNeuS: Neural Surface Reconstruction in Endoscopy Using Illumination Decline. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 2023.
6. H. Itoh, M. Oda, K. Jiang, Y. Mori, M. Misawa, S. Kudo, K. Imai, S. Ito, K. Hotta, K. Mori. Uncertainty meets 3D-spatial feature in colonoscopic polyp-size determination. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 10(3):289–298, 2022, Taylor & Francis.

7. H. Itoh, M. Oda, K. Jiang, Y. Mori, M. Misawa, S. Kudo, K. Imai, S. Ito, K. Hotta, K. Mori. Binary polyp-size classification based on deep-learned spatial information. *IJ Computer Assisted Radiology and Surgery*, 2021.
8. H. Itoh, H. R. Roth, L. Lu, M. Oda, M. Misawa, Y. Mori, S. Kudo, K. Mori. Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*, pages 611–619
9. T. Baltrušaitis, C. Ahuja, L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE T. Pattern Analysis and Machine Intelligence*, 423–443, 2018.
10. M. Misawa, S. Kudo, Y. Mori, K. Hotta, K. Ohtsuka, T. Matsuda, S. Saito, T. Kudo, T. Baba, F. Ishida, H. Itoh, M. Oda, K. Mori. Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointestinal Endoscopy*, 93(4): 960-967.e3, 2021.
11. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*, 2021.
12. R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2020.
13. C. Biffi, G. Antonelli, S. Bernhofer, C. Hassan, D. Hirata, M. Iwatate, A. Maieron, P. Salvagnini, and A. Cherubini. REAL-Colon: A dataset for developing real-world AI applications in colonoscopy. Figshare+. Public Database link *https://doi.org/10.25452/figshare.plus.22202866.v1* , Scientific Data, 11-1-539, 2024. Nature Publishing Group UK London
14. H. Itoh, M. Misawa, Y. Mori, M. Oda, S. Kudo, K. Mori. SUN Colonoscopy Video Database. 2020. URL: http://amed8k.sundatabase.org/
15. B. Sudarevic, P. Sodmann, I. Kafetzis, J. Troya, T. J. Lux, Z. Saßmannshausen, K. Herlod, S. A. Schmidt, M. Brand, K. Sch"ottker, et al. Artificial intelligence-based polyp size measurement in gastrointestinal endoscopy using the auxiliary waterjet as a reference. *Endoscopy*, 55(09):871–876, 2023.
16. K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
17. G. Roffo, S. Melzi, M. Cristani. Infinite Feature Selection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4202–4210, 2015.
18. Y. Tian, G. Maicas, L. Z. C. T. Pu, R. Singh, J. W. Verjans, G. Carneiro. Few-shot Anomaly Detection for Polyp Frames from Colonoscopy. In *MICCAI 2020: 23rd International Conference*, pages 274–284, 2020.
19. R. Yu, A. Li, C.-F. Chen, J.-H. Lai, V. I. Morariu, et Al. NISP: Pruning Networks Using Neuron Importance Score Propagation. In *IEEE CVPR*, 2018.
20. A. Verikas, M. Bacauskiene. Feature selection with neural networks. *Pattern Recognition Letters*, 23(11):1323–1335, 2002.
21. G. Roffo, S. Melzi, “Online feature selection for visual tracking,” in *The British Machine Vision Conference (BMVC)*, 2016, pp. 1–12, GBR.
22. J. Zhong, W. Wang, H. Wu, Z. Wen, J. Qin. PolypSeg: An Efficient Context-Aware Network for Polyp Segmentation from Colonoscopy Videos. In *MICCAI 2020: 23rd International Conference*, pages 285–294, 2020.
23. N. Duffield, C. Lund, M. Thorup. Learn more, sample less: control of volume and variance in network measurement. *Transactions on Information Theory*, 2005.
24. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training recurrent neural networks. International conference on machine learning (ICML), 2013

25. A. Rau, B. Bhattarai, L. Agapito, D. Stoyanov. Task-Guided Domain Gap Reduction for Monocular Depth Prediction in Endoscopy. *MICCAI W. on Data Engineering in Medical Imaging*, pages 111–122, 2023.
26. G. Roffo, S. Melzi, "Feature selection via eigenvector centrality," in *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2016, pp. 1–12, Springer.
27. Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. Advances in neural information processing systems. 2006
28. Deng W, Feng Q, Gao L, Liang F, Lin G. Non-convex learning via replica exchange stochastic gradient MCMC. International Conference on Machine Learning 2020.
29. A. Krizhevsky Learning multiple layers of features from tiny images. (CIFAR-100). Technical report, 2009.
30. Wu K, Zhang J, Peng H, Liu M, Xiao B, Fu J, Yuan L. Tinyvit: Fast pretraining distillation for small vision transformers. In European Conference on Computer Vision (ECCV), 2022 pp. 68-85. Springer Nature Switzerland.
31. Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI), 2022
32. C. Hassan, E. Quintero, J. M. Dumonceau, et al. Post-polypectomy colonoscopy surveillance: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. *Endoscopy*, 45(10):842–851, 2013.
33. M. Ferlitsch, A. Moss, C. Hassan, et al. Colorectal polypectomy and endoscopic mucosal resection (EMR): European Society of Gastrointestinal Endoscopy (ESGE) Clinical Guideline. *Endoscopy*, 49(3):270–297, 2017.
34. N. Gupta, A. Bansal, D. Rao, et al. Prevalence of advanced histological features in diminutive and small colon polyps. *Gastrointestinal Endoscopy*, 1244–1249, 2011.
35. I. Popescu Crainic, R. Djinbachian, D.K. Rex, A. Barkun, A. Shaukat, J. East, C. Hassan, Y. Mori, H. Pohl, A. Rastogi, P. Sharma. Expert endoscopist assessment of colorectal polyp size using virtual scale endoscopy, visual or snare-based estimation: a prospective video-based study. *Scandinavian Journal of Gastroenterology*, 2024.
36. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 3-19).