Video Capsule Endoscopy Classification using Focal Modulation Guided Convolutional Neural Network

Abhishek Srivastava*, Nikhil Kumar Tomar†, Ulas Bagci*, Debesh Jha*
* Machine and Hybrid Intelligence Lab, Department of Radiology, Northwestern University, USA
† School of Computer Science and Informatics, Indira Gandhi National Open University

***Abstract*—Video capsule endoscopy is a hot topic in computer vision and medicine. Deep learning can have a positive impact on the future of video capsule endoscopy technology. It can improve the anomaly detection rate, reduce physicians' time for screening, and aid in real-world clinical analysis. Computer-Aided diagnosis (CADx) classification system for video capsule endoscopy has shown a great promise for further improvement. For example, detection of cancerous polyp and bleeding can lead to swift medical response and improve the survival rate of the patients. To this end, an automated CADx system must have high throughput and decent accuracy. In this study, we propose *FocalConvNet*, a focal modulation network integrated with lightweight convolutional layers for the classification of small bowel anatomical landmarks and luminal findings. FocalConvNet leverages focal modulation to attain global context and allows global-local spatial interactions throughout the forward pass. Moreover, the convolutional block with its intrinsic inductive/learning bias and capacity to extract hierarchical features allows our FocalConvNet to achieve favourable results with high throughput. We compare our FocalConvNet with other state-of-the-art (SOTA) on Kvasir-Capsule, a large-scale VCE dataset with 44,228 frames with 13 classes of different anomalies. We achieved the weighted F1-score, recall and Matthews correlation coefficient (MCC) of 0.6734, 0.6373 and 0.2974, respectively, outperforming SOTA methodologies. Further, we obtained the highest throughput of 148.02 images/second rate to establish the potential of FocalConvNet in a real-time clinical environment. The code of the proposed FocalConvNet is available at https://github.com/NoviceMAn-prog/FocalConvNet.**

***Index Terms*—Video capsule endoscopy, small intestine, deep learning, Kvasir-Capsule, gastrointestinal image classification**

I. Introduction

Wireless capsule endoscopy (WCE) is a technology that allows gastroenterologists to visualize the small bowel (intestine). Video capsule endoscopy (VCE) is useful for differentiating small bowel abnormalities such as small intestine bleeding and inflammatory bowel disease [1]. The capsule-shaped pill (Figure 1) can be swallowed by the patients in the presence of clinical experts without any discomfort [2]. Unlike conventional endoscopy procedures, this procedure investigates the small bowel without pain, sedation and air insufflation. An additional advantage of VCE is that it is non-invasive and easy procedure but plays a crucial role in examination and diagnosis small bowel lesions [1].

The small intestine located inside the gastrointestinal tract (GI) tract performs the powerful function of absorbing nutrients [4]. Ailments caused inside the small intestine can cause grave problems like growth retardation or nutrient deficiencies [4]. A major obstacle in the detection and subsequent treatment of these diseases is the anatomical location of the small intestine. While imaging of the upper GI and large intestine is feasible with endoscopy based methods, the small intestine requires VCE. Figure 1 shows the Olympus EC-S10 endocapsule which was used in Olympus Endocapsule 10 system [5] for capturing videos in the Kvasir-Capsule [3] dataset. VCE is performed by a small capsule equipped with a wide-angle camera. The patient consumes the capsule and the capsule traverses throughout the GI tract and records the video which is later evaluated by a clinical expert. Given the large number of frames generated by the procedure, the subsequent analysis is monotonous and susceptible to human error, which can potentially lead to a high-miss rate [6].

Fig. 1: Olympus EC-S10 endocapsule [3]

CADx systems can save significant time and resources involved in VCE analysis. In this study, our aim is develop an automated VCE deep learning algorithm that can classify VCE frames with high accuracy in real-time so that it can be used in clinical practice. To achieve this, we propose a new neural network architecture, named "*FocalConvNet*", which integrates Convolutional Neural Network (CNN) based modules and Focal modulation [7] to establish a new baseline on Kvasir-Capsule [3].

The main contributions of this work are summarized as follows:

1) We proposed a novel deep learning architecture, named FocalConvNet, to classify anatomical and luminal findings in video frames of capsule endoscopy. We leverage the focal modulation strategy of aggregating multi-scale context for modulating the query. The focal modulation mechanism is integrated with lightweight convolutions,

outperforming several SOTA image classification methods.
2) FocalConvNet not only achieved the best F1-score and MCC in classification on Kvasir-Capsule, but also its lightweight and computationally economical architecture obtained the highest throughput of 148 images per second.
3) To advance the work on long-tailed anatomical, pathological, and mucosal view classification in VCE frames, we provided additional comparisons with SOTA CNN and transformer based classification networks which otherwise lacked comparable studies.

II. Related work

A. Image classification using convolutional neural networks

The introduction of AlexNet [8] led convolutional neural networks to be the most prevalent architectures for nearly all computer vision tasks. Since then tremendous advances have been made with new architectural designs for improving deep neural networks' performance. VGG [9] increased the effective depth in CNNs. ResNet [10] introduced residual connections in CNN. DenseNet [11] proposed densely connected convolutional structures. Multi-scale fusion was introduced in HRNet [12] and further studied in [13], [14]. Spatial and channel-wise attention [15], [16] further enhanced the performance of existing CNNs. Additionally, alterations have been made in convolution operation in depth-wise convolution layer [17] and point-wise convolution [18] layer. Recently, methods like ConvNeXt [19] and ConvMixer [20] have integrated some aspects of vision transformers to further increase the performance of CNNs.

B. Image classification using vision transformers

ViT [21] introduced transformers for vision problems. Since then, this design of architecture has drawn great attention and many incarnations of ViT have been introduced. Swin Transformer used shifting windows for constraining self-attention computation to non-overlapping windows. DeiT [22] relied on a token-based distillation strategy to train data efficient transformers. Further work has been done to reduce the quadratic complexity of the self-attention heads. PoolFormer [23] replaced the attention layer with a pooling operation to achieve favourable results. Focal networks [7] replaced self-attention with modulation of the query by aggregating global and local contexts. Although vision transformers have demonstrated proficiency in learning global representations, CNNs have several desirable properties. Thus, various methodologies blending convolution and self-attention mechanisms have been put forth recently. ConViT [24] used gated positional encoding which combined self-attention with a soft convolutional inductive bias. Although such hybrid networks have been studied in the past, in this work, we combine the efficacy of focal modulation with computationally economical convolutions to introduce a faster and more accurate network for classifying anatomical landmarks, pathological findings, and quality of the mucosal view. Additional work has been done on further reducing architecture complexity and increasing efficiency by; employing sparse attention [25]–[27], Pyramidal designs [28], [29], integration with CNNs [24], [30]. More information on the advancements in vision transformers can be found in [31].

III. Network architecture

The architecture of FocalConvNet consists of three components: initial convolutional stem, FocalConv blocks, and the final linear classifier. FocalConv blocks incorporate alternate convolutional and focal modulation blocks (see Figure 2(b) and Figure 2(d)). Multi-scale context can be gathered and used for modulating the input query through focal modulation blocks. Thus, modelling input dependent global interactions is not as computationally expensive as self-attention strategies utilized by other prominent vision transformers [7]. Keeping with the theme of reducing the computational complexity, our convolutional block uses depth-wise separable convolutions which significantly reduces the required computational resources as compared to standard convolutional layers. Blending the two architectures allows us to leverage global multi-scale context while extracting hierarchical features using convolutional layers and retain the advantageous properties of CNNs (scale and translation invariance, inductive bias) and transformers (increased generalizability).

A. Convolutional Block

Figure 2(b) illustrates the design of our convolutional block. First, the input feature maps are passed through a convolutional layer with a kernel size of 1 (point-wise convolutional layer). Hereafter, a depth-wise convolutional layer with kernel size 3 is used. The ensuing Squeeze and excitation (S&E) block increases the network's representative power by computing the inter-dependencies between channels. Again, a point-wise convolution operation is used, to keep the model parametrically cheap while retaining the power of a standard convolution operator with kernel size 3. The output from this block is fed into the Focal Modulation block.

B. Context Aggregation and Focal Modulation

The input feature map X_f is first operated upon by a linear layer to convert it into a new feature space (see Equation 1).

$$M_0 = Linear(X_f). \tag{1}$$

Let the number of focal levels be n. For the i'th focal level the output M_i is calculated as:

$$M_i = GeLU(DepthConv(M_{i-1})), \tag{2}$$

where $DepthConv$ is a depth-wise convolutional layer followed by $GeLU$ activation layer. Again depth-wise convolutional layer serves in modeling long range interplay. At each layer l, the receptive field can be calculated by $r^l = 1 + \sum_{i=1}^{l}(k^l - 1)$, where k^l denotes the kernel size for layer l. For capturing complete global scale context, average pooling is used to calculate the final $(n+1)$'th feature map, $M_{n+1} = Pool(M_n)$. Hence, the depth-wise convolutional layer is used to obtain hierarchical features, which are used for assembling multi-scale context. Next, a gating mechanism

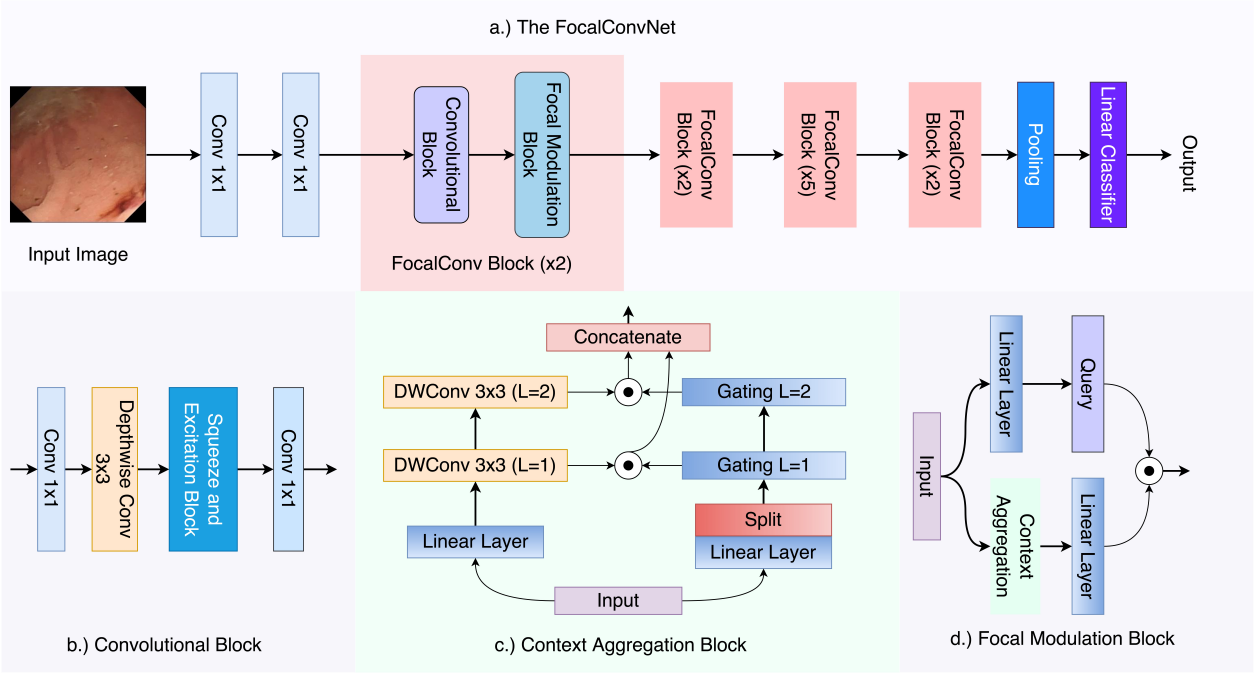

Fig. 2: **The proposed FocalConvNet architectures. a**) The complete FocalConvNet architecture (xK represents that the block in question has been used K times in succession), **b**) The convolutional block used with depth-wise and point-wise convolutions, **c**) Context aggregation sub-module (the figure demonstrates context aggregation when number of focal levels is 2, and can be extended in a similar manner if deeper focal levels are used), **d**) Focal modulation mechanism.

is used to adaptively limit the information propagated by each feature map M_i for each focal level i. A learnable linear layer is used for obtaining gating weights $G = Linear(X_f)$, which has the same dimensions as feature vector M. Finally, M^{out} is obtained by Equation 3

$$M^{out} = \sum_{i=1}^{i=n+1} G_i \odot M_i. \tag{3}$$

Here, $\odot$ represents element-wise multiplication. Hereafter, aggregation is performed across channels using $M^{out} = f(M^{out})$, where f is a linear layer. Subsequent focal modulation is performed by the interaction of query(q) with the modulator M^{out} as $y = q \odot M^{out}$, where y is the output feature vector of the focal modulation block.

C. The FocalConvNet architecture

Initially, the input image is processed by two convolutional layers with a kernel size of 3, where the first and the second layer have a stride of 2 and 1, respectively. Next the extracted features are fed into 4 consecutive sequences of FocalConv blocks (see Figure 2(a)). The aforementioned four sequences consist of 2, 2, 5, and 2 FocalConv blocks, respectively. Each sequence has a focal level of 3 and reduces the spatial dimension of the input feature vector by a factor of 2. Finally, an adaptive average pooling layer and a linear classification layer are used to obtain the final prediction.

IV. Experimental Setup

In this section, we provide details about the dataset, evaluation metrics, implementation details, and data augmentation techniques we used in our experiments.

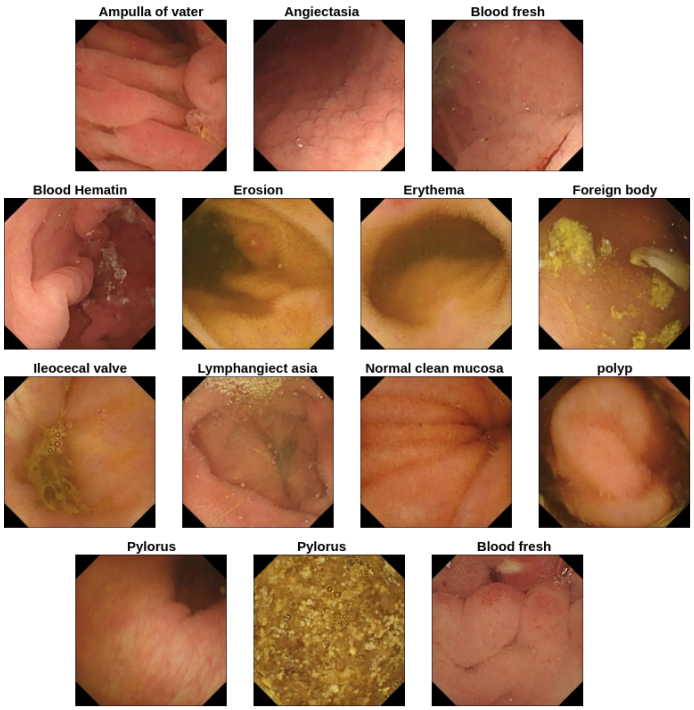

Fig. 3: Example images from anatomy and luminal finding classes from Kvasir-Capsule dataset [32]

A. Dataset

We use Kvasir-Capsule [3], the world's largest video capsule endoscopy dataset for experimentation. The Kvasir-Capsule comprises of 44,228 labelled images from 13 classes of anatomical and luminal findings. Figure 3 shows the examples images from the Kvasir-Capsule dataset. The number of

TABLE I: Result comparison on Kvasir-Capsule [3]

Method	Macro Average			Weighted Average			Accuracy	MCC
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
GMSRF-Net [14]	0.1568	0.1980	0.1575	0.7431	0.6095	0.6636	0.6090	0.2665
ResNet-152 [10]	0.1563	0.2049	0.1463	0.7295	0.4886	0.5675	0.4886	0.1979
DenseNet-169 [11]	0.1884	0.2262	0.1483	0.7141	0.5540	0.6083	0.5540	0.2041
ConvMix-768/32 [20]	0.1426	0.1779	0.1188	0.7464	0.3475	0.4428	0.3475	0.1570
ConvMix-1536/20 [20]	0.1722	0.2275	0.1697	0.7431	0.6021	0.6524	0.6021	0.2717
EfficientNetV2-S [33]	0.1623	0.2424	0.1686	0.7563	0.5588	0.6312	0.5588	0.2615
EfficientNetV2-M [33]	0.1558	0.2102	0.1456	0.7347	0.5887	0.5199	0.5199	0.2267
ConvNeXt-S [19]	0.1177	0.2310	0.1012	0.7277	0.4349	0.5173	0.4349	0.1758
ConvNeXt-B [19]	0.1311	0.1169	0.1108	0.7276	0.3917	0.4965	0.3918	0.1387
Swin-S [26]	0.1538	0.2388	0.1525	0.7390	0.5800	0.6334	0.5800	0.2565
Swin-B [26]	0.1496	0.2310	0.1525	0.7134	0.5905	0.6355	0.5905	0.2288
ConViT-S [24]	0.1765	0.2182	0.1689	**0.7673**	0.5610	0.6312	0.5610	0.2769
ConViT-B [24]	0.1769	0.2534	0.1700	0.7406	0.5541	0.6160	0.5541	0.2443
Focal-S [7]	0.1403	0.1919	0.1344	0.7352	0.4883	0.5690	0.4883	0.2060
Focal-B [7]	0.1394	0.1869	0.1368	0.7300	0.5110	0.5873	0.5110	0.2092
FocalConvNet(Ours)	**0.2438**	**0.2745**	**0.2178**	0.7557	**0.6373**	**0.6734**	**0.6373**	**0.2964**

images per class is as followed; Normal mucosa - 34,606; Reduced Mucosal View - 2399; Pylorus - 1520; Polyp - 64; Lymphoid Hyperplasia - 592; Ileo-Cecal valve - 1417; Hematin - 12; Foreign Bodies - 776; Erythematous - 238; Erosion - 438; Blood - 446; Angiectasia - 866; Ulcer - 854. We observe that the number of images in class "Normal Mucosa" is significantly large as compared to any other class; hence, the dataset is heavily imbalanced. As done by the authors in [3], we remove the "Polyp" and "Hematin" classes as the number of images is very scarce in these classes.

B. Implementation details

We used 23,061 images for training and 24,092 for testing. Random horizontal, vertical flipping and random rotation were used to augment data. Images were resized to 224×224 before being fed into the models. The implementation of our proposed FocalConvNet was done using the PyTorch framework. SGD optimizer with a momentum of 0.9 and a learning rate of 0.001 were used during training. We used author-released source code for training and each method was trained from scratch. Throughput was calculated on a single Tesla V100 GPU with a batch size of 6. We used weighted categorical cross-entropy as our loss function. Each experiment was performed on NVIDIA DGX-2 machine that uses NVIDIA V100 Tensor core GPUs.

C. Evaluation metrics

For evaluation purposes, we have chosen standard computer vision metrics such as F1-score, precision, recall, accuracy, throughput, and MCC. MCC and F1-score are the most preferred metric for the class imbalance problem. Detailed information about the evaluation metrics can be found in [32].

V. Result and Discussion

We compared our proposed FocalConvNet with other SOTA baselines on the Kvasir-Capsule dataset as follows. Table I shows the quantitative comparison where both the "weighted" and "macro" averaging strategies were used while calculating F1-score, precision and recall. Additionally, we calculated the MCC achieved by each method. From Table I, we observed that FocalConvNet attained a 0.98% and 2.78% improvement in F1-score and accuracy, respectively, over second-best performer GMSRF-Net. The MCC reported by FocalConvNet was a 1.95% increment over best performing ConViT-S. Additionally, we noted the highest precision, recall, and F1-score of 0.2438, 0.2745, and 0.2178, respectively, when macro averaging was used. Consequently, we inferred that the multi-scale context guided visual modelling coupled with the inherent inductive bias of convolutions within our FocalConvNet was advantageous and powerful while extracting discriminative features.

TABLE II: Comparison of FocalConvNet with baselines in terms of computational requirement (M denotes million) and throughput(number of images processed per second).

Method	Paramaters	Year	GFLOPs	Throughput
GMSRF-Net [14]	55.08 M	2021	160.88	7.60
ResNet-152 [10]	58.17 M	2016	11.58	53.16
DenseNet-169 [11]	26.5 M	2017	7.82	130.79
ConvMix-768/32 [20]	20.35 M	2022	20.88	49.34
ConvMix-1536/20 [20]	50.11 M	2022	51.36	19.20
EfficientNetV2-S [33]	22.17 M	2021	2.97	65.23
EfficientNetV2-M [33]	51.11 M	2021	6.24	59.17
ConvNeXt-S [19]	49.46 M	2022	8.70	69.53
ConvNeXt-BS [19]	87.58 M	2022	15.38	50.36
Swin-S [26]	48.85 M	2021	8.52	102.07
Swin-B [26]	86.75 M	2021	15.14	83.67
ConViT-S [24]	27.35 M	2021	5.35	136.66
ConViT-B [24]	85.78 M	2021	16.8	45.56
Focal-S [7]	27.89 M	2022	8.70	134.98
Focal-B [7]	49.13 M	2022	15.40	123.27
FocalConvNet(Ours)	34.66 M	2022	5.23	**148.02**

Table II compares the parameters, one billion floating-point operations per second (GFLOPs), and Throughput (images/second) of the proposed method with other baselines. We noted that FocalConvNet along with obtaining the highest weighted average F1-score, MCC and accuracy, achieved the highest throughput of 148.02 despite having greater computational complexity than Focal-S, ConViT-S, EfficientNetV2-S, ConvMix-768/32 and DenseNet-169. Even though Focal-S and Focal-B used focal modulation as the key element in their architecture, it remained devoid of aggregation of local

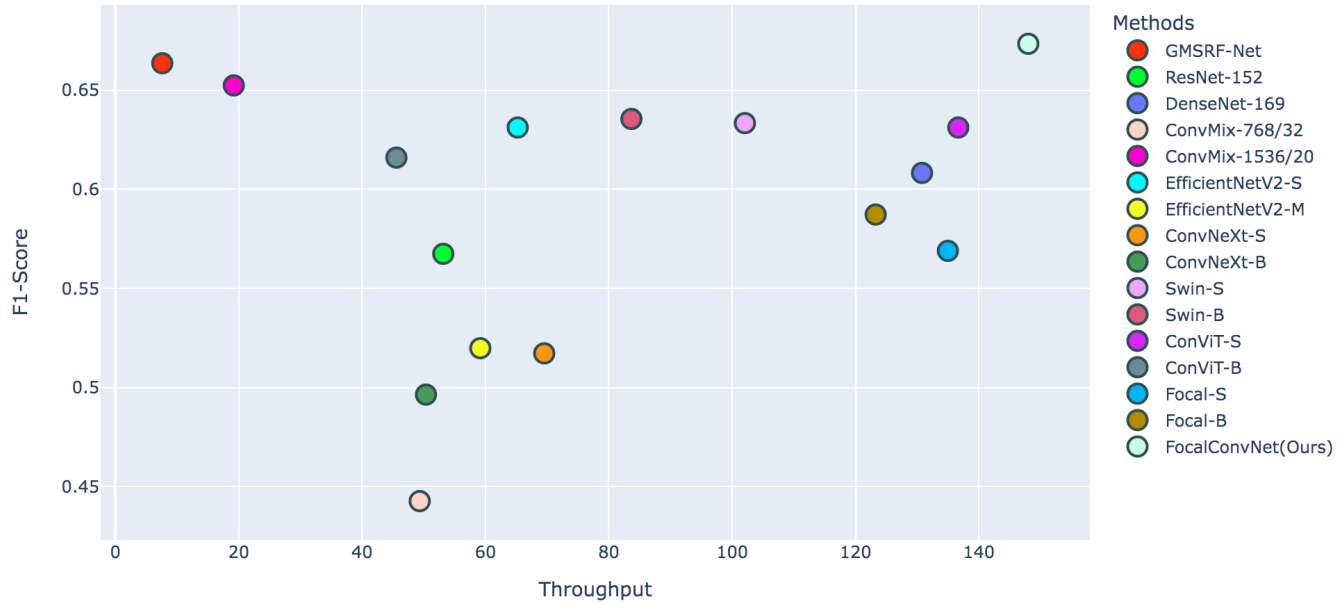

Fig. 4: **Plot of F1-score vs throughput for each method**. We demonstrate that FocalConvNet is capable of achieving the best weighted F1-score while maintaining a high throughput.

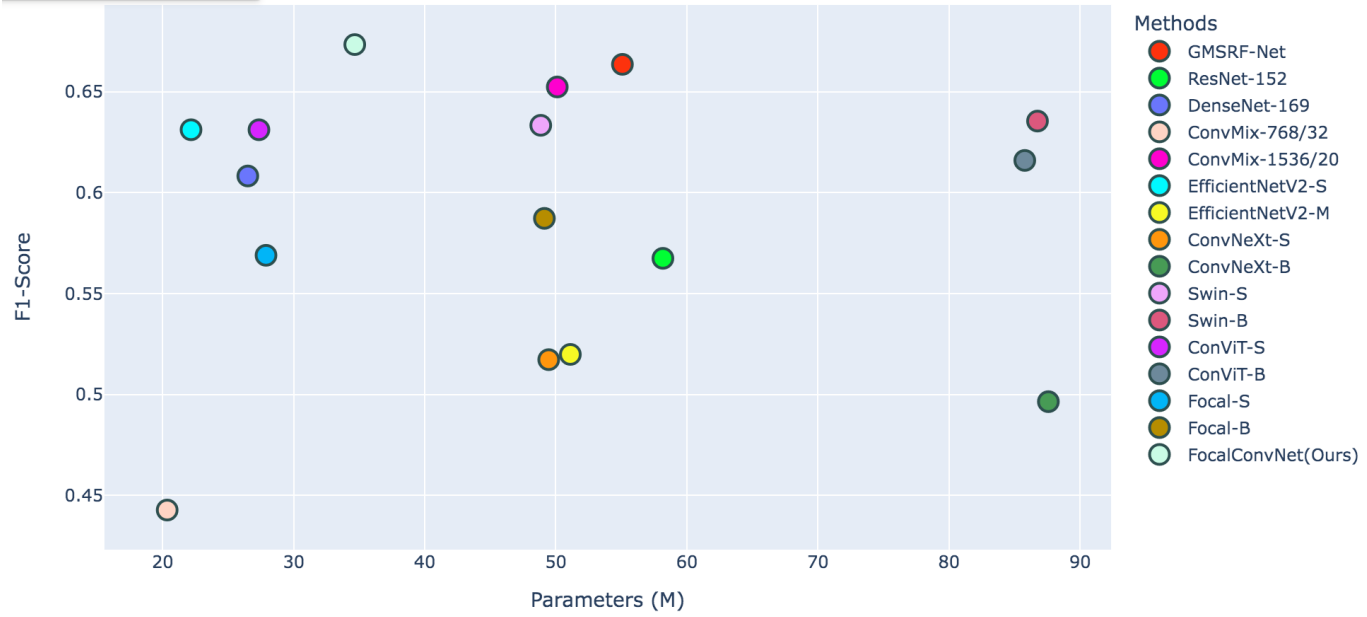

Fig. 5: **Plot of F1-score vs parameters for each method**. Here, we can observe the trade-off between number of parameters and performance for all methods.

features which can otherwise be achieved by convolutional layers. Consequently, it gets outperformed by several other CNN and transformer-based methods. Both the variants of Swin and ConViT obtained significant performance underscoring the power of vision transformers to effectively learn global and local interactions. Figure 4 illustrates the F1-score vs throughput observed by each method. Here, we ascertained that our proposed method was superior in performance when both F1-score and inference time was considered. An identical trend was followed when recall and MCC were plotted against throughput for all methods. GMSRF-Net utilizing its global multi-scale fusion mechanism learned strong representations boosted by the global context and reports comparable results. Nevertheless, its high computation requirements (160.88 GFLOPs) and lowest throughput (7.60 images/second) were not favourable in a real-time setting.

This property of the FocalConvNet can be attributed to focal modulation and the lightweight convolutional block. Focal modulation is proficient in learning visual token interactions while serving as a light-weight replacement for self-attention. The convolutional block leverages 1×1 convolutions and depth-wise convolutions to serve as a proxy for computationally heavy 3×3 convolutions. Thus, in medical image classification where data tends to be scarce or imbalanced in nature, our FocalConvNet can serve as an effective and efficient baseline in future. Moreover, in Figure 5 it can be seen

that in the performance vs complexity graph, FocalConvNet shows a decent trade-off.

VI. Conclusion

In this study, we propose a novel lightweight and swift medical classification architecture for real-time anatomical and luminal findings (pathological and mucosal view) classification in video capsule endoscopy. The proposed deep network, FocalConvNet, leverages the learning bias in convolutions and mixes it with the global and local representation learning power of focal modulation to give favourable performance on Kvasir-Capsule. Combination of focal modulation and lightweight convolutions enables the FocalConvNet to not only outperform other SOTA baselines, but also report the highest throughput. We showed that our proposed method can be used in real-time analysis of video capsule endoscopy. Thus, reducing the effort required for manual inspection. We also use several SOTA baselines and benchmark them on Kvasir-Capsule to streamline further research in this area. In the future work, we plan to leverage generative models to abate the class imbalance and further increase the performance.

Acknowledgement: This project is supported by the NIH funding: R01-CA246704 and R01-CA240639.

References

[1] S. Y. Lee, J. Y. Lee, Y. J. Lee, and K. S. Park, "Natural elimination of a video capsule after retention for 1 year in a patient with small bowel crohn disease: A case report," *Medicine*, vol. 98, no. 43, 2019.

[2] S. Suman, A. S. Malik, K. Pogorelov, M. Riegler, S. H. Ho, I. Hilmi, K. L. Goh *et al.*, "Detection and classification of bleeding region in wce images using color feature," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing*, 2017, p. 17.

[3] P. H. Smedsrud, V. Thambawita, S. A. Hicks, H. Gjestang, O. O. Nedrejord, E. Næss, H. Borgli, D. Jha, T. J. D. Berstad, S. L. Eskeland *et al.*, "Kvasir-capsule, a video capsule endoscopy dataset," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[4] G.-V. Meerveld, A. C. Johnson, D. Grundy *et al.*, "Gastrointestinal physiology and function," *Gastrointestinal pharmacology*, pp. 1–16, 2017.

[5] Olympus, "The endocapsule 10 system," *Olympus homepage*, https://www.olympus-europa.com/medical/en/Products-and-Solutions/Products/Product/ENDOCAPSULE-10-System.html, 2013.

[6] M. F. Kaminski, J. Regula, E. Kraszewska, M. Polkowski, U. Wojciechowska, J. Didkowska, M. Zwierko, M. Rupinski, M. P. Nowacki, and E. Butruk, "Quality indicators for colonoscopy and the risk of interval cancer," *New England Journal of Medicine*, vol. 362, no. 19, pp. 1795–1803, 2010.

[7] J. Yang, C. Li, and J. Gao, "Focal modulation networks," *arXiv preprint arXiv:2203.11926*, 2022.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[12] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[13] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "MSRF-Net: A Multi-Scale Residual Fusion Network for Biomedical Image Segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2252–2263, 2022.

[14] A. Srivastava, S. Chanda, D. Jha, U. Pal, and S. Ali, "GMSRF-Net: An improved generalizability with global multi-scale residual fusion network for polyp segmentation," in *Proceedings of the International conference on pattern recognition*, 2022.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] 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, pp. 3–19.

[17] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[18] B.-S. Hua, M.-K. Tran, and S.-K. Yeung, "Pointwise convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 984–993.

[19] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.

[20] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357.

[23] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," *arXiv preprint arXiv:2111.11418*, 2021.

[24] S. d'Ascoli, H. Touvron, M. L. Leavitt, A. S. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 2286–2296.

[25] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6836–6846.

[26] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," *arXiv preprint arXiv:2107.00652*, 2021.

[27] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," *Advances in neural information processing systems*, vol. 34, 2021.

[28] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[29] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6824–6835.

[30] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[31] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (CSUR)*, 2021.

[32] D. Jha, "Machine learning-based classification, detection, and segmentation of medical images," *PhD thesis*, 2022.

[33] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 10 096–10 106.