Class-aware data augmentation by GAN specialisation to improve endoscopic images classification

Cyprien Plateau–Holleville, Yannick Benezeth
Univ. Bourgogne Franche-Comté, ImViA EA7535, France
yannick.benezeth@u-bourgogne.fr

***Abstract*—An expert eye is often needed to correctly identify mucosal lesions within endoscopic images. Hence, computer-aided diagnosis systems could decrease the need for highly specialized senior endoscopists and the effect of medical desertification. Moreover, they can significantly impact the latest endoscopic classification challenges such as the Inflammatory Bowel Disease (IBD) gradation. Most of the existing methods are based on deep learning algorithms. However, it is well known that these techniques suffer from the lack of data and/or class imbalance which can be lowered by using augmentation strategies thanks to synthetic generations. Late GAN framework progress made available accurate and production-ready artificial image generation that can be harnessed to extend training sets. It requires, however, to deal with the unsupervised nature of those networks to produce class-aware artificial images. In this article, we present our work to extend two datasets through a class-aware GAN-based augmentation strategy with the help of the state-of-the-art framework StyleGAN2-ADA and fine-tuning. We especially focused our efforts on endoscopic and IBD datasets to improve the classification and gradation of these images.**

I. Introduction

The detection and gradation of digestive mucosal lesions require the analysis of endoscopic images by an expert gastrointestinal pathologist. However, the availability of medical practitioners is highly related to geographical location. Inequalities in access to care due to the trained professional unavailability can be offset with the help of Computer-Aided Diagnosis (CAD). These systems offer adequate accuracy levels, which might even outperform the expert eye [1] and be a decision support for non-expert pathologists [2]. Their development can then be an interesting extension of the available medical equipment to facilitate patient recovery.

Both Inflammatory Bowel Disease (IBD) forms, Crohn's Disease (CD) and Ulcerative Colitis (UC), are chronic diseases that mainly cause rectum and colon inflammation. Currently, they represent one of the main classification challenges for automatic systems [3]. Their graduation is commonly performed with state-of-the-art metrics such as the Mayo Score for UC [4] and needs a solid background with continuous training to acquire a good accuracy level. The use of CAD systems might therefore help to offer stable results between patients' specific cases and professional curriculum. Finally, the search for reproducibility in pathological diagnosis is still an active medical research topic [5]. Regarding their technical and mathematical characteristics, CAD systems could then be able to bring more stability to the classification than expert pathologists.

Deep-learning-based solutions, such as Convolutional Neural Network (CNN), have been successful in image classification [6] and offer mandatory robustness for medical use as endoscopic data analysis [7, 1]. However, these algorithms need a substantial amount of data to provide satisfactory performance. This requirement might be hard to meet regarding the scarcity of medical data which can involve a general lack of data and imbalanced classes. The use of augmentation strategies helps to lower the deficit of images in a deep-learning-based project [8] and enhances the size and quality of training datasets. Usually, basic transformations are performed, such as translation and rotation, while preserving the semantical information. This kind of modification can fail to ease the generalization of the training data due to a lack of variation. Generative Adversarial Networks (GAN) have made great progress since they were introduced by [9] and enabled close to photorealism synthesis [10, 11]. The use of artificially generated images can help to increase the sample number of training datasets and ease the training of feature detectors. In this work, we use late progress from GAN research with StyleGAN2-ADA [12, 13] and transfer learning for class-specific artificial image generation [14] to extend endoscopic datasets and improve the accuracy of state-of-the-art classification systems. As we considered the IBD classification as one of the current challenges in endoscopic imaging, we validate our method on a specific UC dataset. Finally, we make the code of this project available*.

The first section of this article focuses on data augmentation and state-of-the-art GAN frameworks. The second presents our image augmentation strategy, followed by the experimental results. Finally, the conclusion summarises the method, its results, and potential future works.

II. State of the art

a) Data augmentation: The artificial extension of the training set has been widely used to improve the accuracy of deep-learning-based systems [8, 15]. Most algorithms provided within deep learning frameworks for this purpose are basic

*Project web page: https://github.com/PlathC/GanBasedAugmentation

image transformations that decrease the risk of positional overfitting by moving the semantic information throughout the image. These image transformations are helpful during neural network training to emphasize the meaningful information in the data. However, this kind of augmentation is limited and can fail to produce the necessary diversity to achieve its goal.

b) Neural augmentation: The use of neural-based solutions for data augmentation can improve classification performance [15] to extend imbalanced or small training sets. This subfield aims to propose data augmentation strategies based on a neural network to improve the training of another learner network. This is performed by producing data to ease the understanding of the student network. One of the main issues faced for its application is to provide sufficient knowledge of the source data to the trainer network based on innovative methods regarding the application context.

c) Generative Adversarial Networks: GAN [9] are neural network systems known for the quality of their image generation. They are trained during a min-max game between a generator that aims to fool a classifier network by producing realistic images from a random noise vector. On the other hand, the second network, a classifier, aims to predict the real nature of its inputs. However, this type of framework originally requires special training conditions and a large amount of data to converge. The use of novel and innovative architecture reduced these problems and helped to democratize this technology [16] as well as the latest breakthroughs in GAN research and synthetic generation [10]. StyleGAN's [11] arrival made available style manipulation which contributes a lot to ease the disentanglement of these systems and to preserve the quality of the produced images or stabilizing the training. StyleGAN2 [12] is an improvement that fixes generation artifacts and decreases computation costs via architecture modifications and simplifications. However, the quality of the generation of these technologies is still linked to the amount of data given during the training phase. This need can be lowered by transfer-learning [17] and specialization of the system on target data. Finally, standard data augmentation strategies have been adapted to fit with GAN needs based on adaptive differentiable algorithms targeting the discriminator and providing backpropagation through these augmentation computations [13]. Therefore, this work helped to democratize the use of GAN on a low amount of data. Based on these features, GAN can be harnessed for neural augmentation. It has moreover already been established in the state of the art for various purposes such as dataset equalization [18] and medical imaging [19, 20, 21, 22].

d) Fréchet Inception Distance: The evaluation of GAN performance has remained a challenge until the introduction of the Fréchet Inception Distance (FID) [23]. This metric intends to measure the realism of synthetic data and has been widely used in GAN research. It is given by:

$$\mathrm{FID}\,(X,Y) = \|\mu_X - \mu_Y\|^2 + \mathrm{Tr}\left(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}\right). \quad (1)$$

This formula aims to provide the distance between the original distribution X and the artificial one Y by comparing means μ and variances based on the covariance matrices traces $\mathrm{Tr}\,(\Sigma)$ of an Inception V3 [24] deep layer activation from both distribution samples. As a distance, the lesser the result of the FID is, the closer are the distributions and the best is the image quality.

e) Conditional generation: To properly extend a training set, the synthetic generation needs to comply with differences across classes within the dataset distribution. GANs are mostly unsupervised systems that require strenuous efforts to set up a class-aware generation. Various methods try to constrain the network's generation to produce specific features [20]. Other methods use latent space exploration thanks to backpropagation or principal component analysis [25, 26] and allow precise control of the generation based on the study of the GAN representation. However, these features require individual domain adaptation and human intervention to acquire semantic accurate results without "leaking" between class boundaries. Transfer learning has finally demonstrated its ability to limit GAN's latent space based on training specialization in a context of limited data [27, 14]. Nevertheless, this last strategy involves the creation and the training of individual weights groups dedicated to each target class.

The data augmentation system we present aims to provide realistic synthesis in a multi-class context of limited data. Hence, we chose StyleGAN2 [12] as the basis of the generation system for its state-of-the-art artificial generation. In addition, the use of the Adaptive Discriminator Augmentation (ADA) [13] is perfectly suited to ease the GAN training with limited data. Finally, the specialization training by transfer learning [14, 28] is one of the most straightforward strategies to restrict the GAN generation space to provide a class-aware synthesis.

III. Method

The method developed in this work is based on the strategy presented by [14] which proposes to extend the training set thanks to transfer-learning for class-aware generations. This specialization aims to lower the need for per-class large amounts of data to get good convergence performance with conditional synthesis based on adversarial training. Moreover, this type of GAN transfer learning has already been demonstrated in medical imaging for other purposes [28]. However, whereas [14] used a dedicated regularization function, we chose to use the freezing of the discriminator higher-level layers [17] to benefit from its balance between simplicity, accessibility, and state-of-the-art effectiveness.

Figure 1 illustrates our method. We first fully pretrain the network on an extensive unlabeled database which aims to acquire basic endoscopic artificial generation knowledge. The trained weights are then reused during a fine-tuning step by freezing the first higher-level layers of the discriminator to take advantage of its feature extraction capabilities and to fine-tune its classification part on a class-specific subset [17]. The goal is to specialize the generator to the target class while using the general knowledge acquired on many samples.

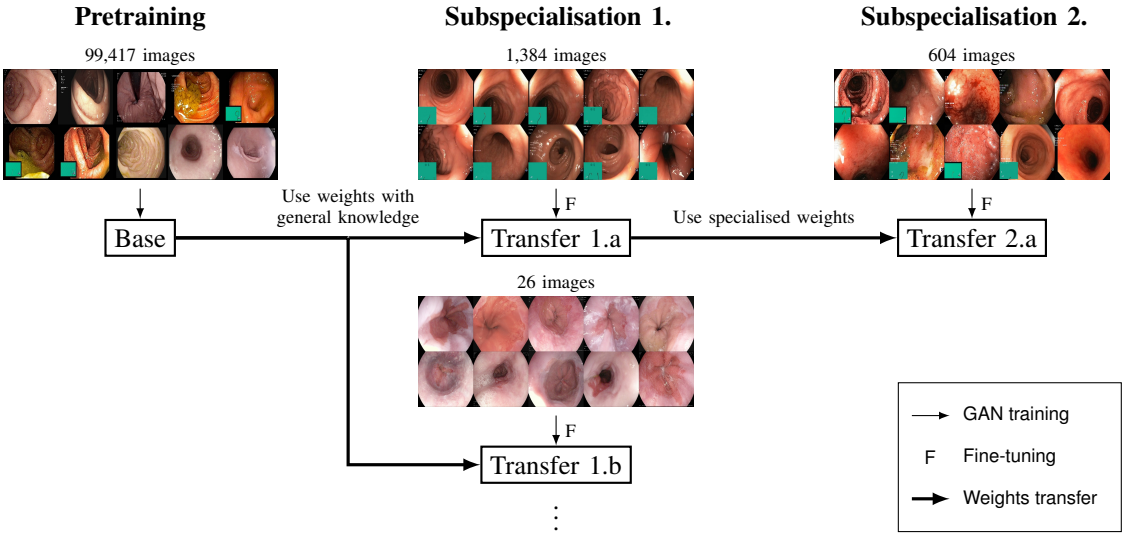


Fig. 1: The transfer learning system used to generate class-aware images.

This specialization can be repeated as much as necessary on more specific data. To obtain good results, a funnel-shaped specialization is used to gradually transfer information to the network and improve the generation quality [14]. This step-wise strategy involves selecting the training order based on class properties to provide an iterative process during the learning of the network.

The GAN architectural choice is a crucial step that may affect the needs of specific experimentation, hyperparameter tuning, training time, and/or image quality. We chose the StyleGAN2 [12] framework for its convergence performance, its stability during training, and its availability. The additional use of adaptive discriminator data augmentation [13] suited particularly well with the goal of this project since it provides significant improvements during the training on limited data.

To finally extend the training set such that the classifier better understands the feature we are aiming to classify, one prime setting is to select the classes that need to be extended and in which proportions. We decided to stick our experiments to a uniform augmentation to limit their number.

IV. Experimentation

This section describes how we used the GAN-based augmentation method on gastrointestinal datasets, validate it on IBD gradation, and the results of our experiments.

A. Data

At first, research efforts on gastrointestinal disease classification were focused on polyps which are currently highly represented in existing datasets. To handle other diseases and work on more challenging classification issues, appropriate datasets were created to provide training data as a basis for new method developments. These are, however, still unusual, and only a few contain images of new challenging topics such as IBD. We then collected all candidates for this application related to our need for endoscopic images in Table I.

Hyper-Kvasir [29] presents the main advantages of aggregating large panels of gastrointestinal data in a broader way than CAD-CAP [30] and Crohn-IPI [31]. Its wide number of classes also enables to test our method in different configurations with varying sample numbers or with diverse feature panels which can also contain characteristics not produced by diseases (tool appearing on image or green thumbnail representing the topographic view of the colon). The unlabeled 90,000 images of the dataset are moreover highly attractive as a basis of the transfer-learning pipeline.

To perform the validation of our method, an additional *Hyper-Kvasir* [29] subset, described in Table II, has been created to focus our experiments on UC and to provide an alternative study with fewer classes than the base dataset. This partial dataset aims to explore our method on the classification challenge that IBD is currently representing. We referred to it as the *Custom-UC* set. Moreover, it lays the foundation of an experimental protocol regarding its limited number of samples.

Hyper-Kvasir's [29] UC classes are split into ascending levels (0 to 3) by the Mayo Score [4] and defined by experts evaluation which introduces intermediate levels due to confusing cases. We decided to merge classes based on this score to create a bias-free experimental protocol and provide a more production-like environment [4, 32]. Since its first two levels indicate an inactive (grade 0) and a light inflammation (grade 1), we combined those into a single class called non-pathological. We added images from the Boston Bowel Preparation Scale (BBPS) [33] classes of levels 2 and 3 to this class to augment its number of images. Indeed, the BBPS score evaluates cleanliness ascendingly (0 to 3), 2 and 3 being the ones where the bowel mucosa has been flushed and can be correctly observed. Even if *Hyper-Kvasir* [29] does not provide a dedicated "Healthy" class, the authors define BBPS classes as non-pathological findings which can then be leveraged as normal bowel images. The remaining UC classes, except grade 1-2, were added to the pathological class since they describe moderate and severe

Name	Diseases	Modality	Size-Balance	Availability
Hyper-Kvasir [29]	Polyps, Barrett's esophagus, Ulcerative colitis...	White light	110.079 images, 10 662 labeled	Open academic
CAD-CAP [30]	Polyps, Vascular lesions Ulcerative lesions...	White light, WCE	24,824 images 4,824 pathological 20,000 non-pathological	By request
Crohn-IPI [31]	Crohn's disease	WCE	3498 images 40% pathological 60% non-pathological	By request

TABLE I: Considered dataset for IBD classification

inflammation. Indeed, we chose to avoid using the UC grade 1-2 for the *Custom-UC* set that could not have been classified as healthy or pathological and could have introduced a bias.

Non-pathological		Pathological	
Class name	**Image number**	**Class name**	**Image number**
UC 0-1	35	UC-2	133
UC 1	201	UC-2-3	28
BBPS2-3	1148	UC-3	443
Total	1,384	**Total**	604

TABLE II: Distribution of non-pathological and pathological classes of the *Custom-UC* set and their source classes in *Hyper-Kvasir* [29]

Since [29] provided fixed splitting of the dataset for reproducibility purposes, we chose to use the first split (numbered 0) as the training set, and the second split (numbered 1) as the validation set. These two parts use a ratio of 50/50, which divides by two the number of available images within the training set and strengthen the need for data augmentation.

B. Experimental settings

a) GAN settings: Each GAN training has been performed until the FID stopped improving. We used a batch size of 16, an image size of 256^2, enabled dataset mirroring augmentation, and kept all default remaining StyleGAN2-ADA [12, 13] settings. All experimentations were performed with Nvidia P100 and V100 depending on availability with 16 Gb of VRAM.

b) Classifier architectures: Classifier architectures have been selected based on the work presented by [29]. Some of their published experimentations were carried out through ResNet [34] and DenseNet [35]. These well-known CNN architectures are time-tested in the current context and allow state-of-the-art accuracy levels especially by including a pretraining phase on ImageNet [36] to acquire general features detection capabilities. We selected for our experiments ResNet-50 and DenseNet-161 networks rather than ResNet-152. Indeed, the latter has a higher number of parameters than ResNet-50 while having poorer performance than DenseNet-161. Moreover, the Resnet-50 low number of parameters could enable lightweight evaluation and training.

c) Classifiers settings: Training of both ResNet-50 and DenseNet-161 were performed on same hardware configurations and with same image size as generative networks, respectively configured with batch size of 64 and 128. All experimentations were executed based on a Stochastic Gradient Descent with a momentum of 0.9, restricted to 30 epochs for time optimization purposes, include standard data augmentation based on basic image transformations (color jittering, specific rotation $\{0°, 90°, 180°, 270°\}$, random horizontal flips, translations, shears, and crops), and use cyclic learning rate scheduling starting at $1e^{-3}$ with an amplitude of 0.1.

d) Augmentation strategy: We decided to extend the *Custom UC* dataset uniformly while we selected in-tension classes from *Hyper-Kvasir* by analyzing the baseline's prediction results. Indeed, we chose to focus our augmentation strategy for the *Hyper-Kvasir* study on classes that needed it the most and then avoid training one GAN for each of its 23 classes, which would have been computationally and time-consuming. The latter was conducted by taking into account by-class results obtained from raw training, presented in Figure 2, and the number of samples within the relevant classes. We then chose to extend only classes with a few samples as Ileum or Hemorrhoids, and the ones with low performing results as Esophagitis A or Ulcerative Colitis Grade 1 and 3.

C. Results

Experimental results presented in this section contain the evaluation of artificial image quality, and the ability to improve classifier performance based on state-of-the-art metrics. We chose to follow the metrics choice of [29] to ease the comparison with their results and to benefit from the capacity of these indicators to evaluate performance in a multi-class context.

a) Image synthesis: Table III presents the training order of each training setting and its results compared to real randomly selected samples. Artificial results present good visual realism and distribution spread confirmed by the FID score level on the *Custom-UC* classes, which seems to contain a sufficient amount of samples for this augmentation method. These results strengthen the choice of StyleGAN2-ADA [12, 13] with fine-tuning [17] to fit with the needs of this project to provide a quality artificial conditional generation. The augmentation of *Hyper-Kvasir* [29] classes containing fewer samples than our custom subset appears to get poorer results based on the FID score even if class features can be properly restored, as we can see with stage 3 of the Ulcerative Colitis class. Rendered mucosal views contain characteristic inflammatory textures that display the GAN ability to catch the disease attributes but might

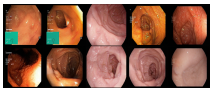
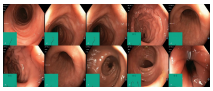
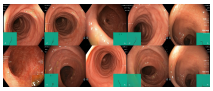
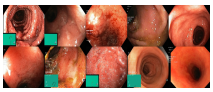
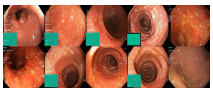
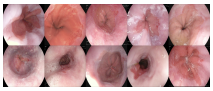
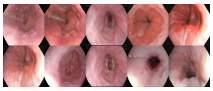
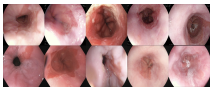
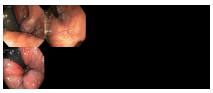
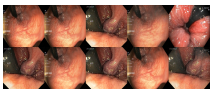
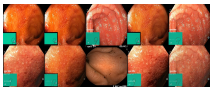
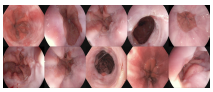
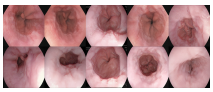
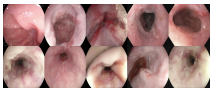
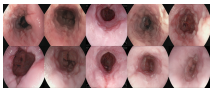
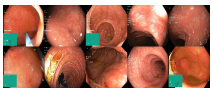
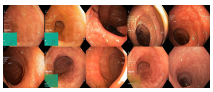
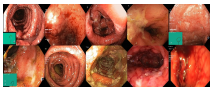
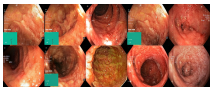
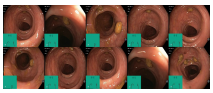
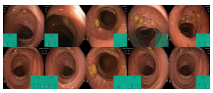
	Real	Synthetic	Starting weights	Training image number	Best FID↓
Unlabeled			None	99,417	18.32
Non Pathological (*Custom-UC*)			Unlabeled	691	43.55
Pathological (*Custom-UC*)			Non Pathological (*Custom-UC*)	301	55.58
Barretts Short Segment			Unlabeled	26	96.75
Barretts Long segment			Unlabeled	20	118.66
Hemorrhoid			Unlabeled	3	97.68
Ileum			Unlabeled	4	65.41
Esophagitis A			Unlabeled	201	51.67
Esophagitis B-D			Unlabeled	130	71.97
Ulcerative Colitis Grade 1			Unlabeled	100	83.92
Ulcerative Colitis Grade 3			Unlabeled	66	111.21
Impacted Stools			Unlabeled	65	76.28

TABLE III: [Best viewed in electronic version.] GAN training settings and randomly selected generated images compared to the original distribution. (↓: Lower is better)

lack variation and fail to map the original distribution correctly. The learning procedure on the few provided hemorrhoid and ileum images strongly overfitted as displayed by the distribution returned by the generators that could not produce more variation. Slight contrasts can still be observed within ileum generations which are similar to basic mixing between training set images and seems insufficient to produce more diversity. Finally, it seems that the image synthesis of classes with only a few samples is badly evaluated by the FID. This leads to a better score for these samples than classes with visually less overfitted production. This may be produced by the bias of the metric introduced by the very small sample number [37].

b) Dataset augmentation: Table IV displays the protocol baseline compared to the best performing augmented configurations. We present the best-augmented results after an iterative experiment to find peaks by adding synthetic images to the training set. It then demonstrates that even if networks have similar augmentation needs, training and architectural choices might lead to different results with our augmentation system.

Results on full *Hyper-Kvasir* [29] demonstrate the method capabilities, especially on non-pretrained networks. Pretrained models manifest more limited improvements on micro average metrics than on macro average, demonstrating the difficulty of extending classes with only a few samples. The chosen strategy can properly generate class samples by reducing the need for a large amount of data to fit with this project's goals but seems to require at least tens of samples. Minor improvements can be observed in the confusion matrices in Figures 2 and 3. Even if several classes are better classified such as Barret's Short Segment, other classes, such as Z-Line, get confusing results after the augmentation, which moderates the improvement.

Training configurations on the *Custom-UC* set validate the

Configuration	Architecture	Pretrained	Artifical addition	Macro Average			Micro Average			
				Precision ↑	Recall ↑	F1 ↑	Precision ↑	Recall ↑	F1 ↑	MCC ↑
Hyper-Kvasir [29]	ResNet-50 [34]	No	-	0.491	0.455	0.461	0.771	0.771	0.771	0.751
			600	**0.521**	**0.485**	**0.491**	**0.786**	**0.786**	**0.786**	**0.768**
		ImageNet	-	0.577	0.593	0.584	0.895	0.895	0.895	0.886
			600	**0.609**	**0.601**	**0.596**	**0.898**	**0.898**	**0.898**	**0.889**
	DenseNet-161 [35]	No	-	0.533	0.494	0.500	0.806	0.806	0.806	0.791
			500	**0.551**	**0.520**	**0.525**	**0.842**	**0.842**	**0.842**	**0.829**
		ImageNet	-	0.602	0.601	0.596	0.902	0.902	0.902	0.894
			500	**0.617**	**0.613**	**0.613**	**0.907**	**0.907**	**0.907**	**0.899**
Custom-UC	ResNet-50 [34]	No	-	0.804	0.815	0.809	0.835	0.835	0.835	0.619
			3000	**0.857**	**0.901**	**0.871**	**0.883**	**0.883**	**0.883**	**0.757**
		ImageNet	-	0.919	0.916	0.917	0.930	0.930	0.930	0.835
			2500	**0.924**	**0.932**	**0.928**	**0.938**	**0.938**	**0.938**	**0.857**
	DenseNet-161 [35]	No	-	0.846	0.888	0.860	0.873	0.873	0.873	0.734
			3000	**0.857**	**0.898**	**0.870**	**0.883**	**0.883**	**0.883**	**0.754**
		ImageNet	-	0.923	0.925	0.924	0.935	0.935	0.935	0.848
			3000	**0.930**	**0.945**	**0.937**	**0.945**	**0.945**	**0.945**	**0.875**

TABLE IV: Comparison of raw training metrics and augmented training metrics based on our method (↑: Higher is better)

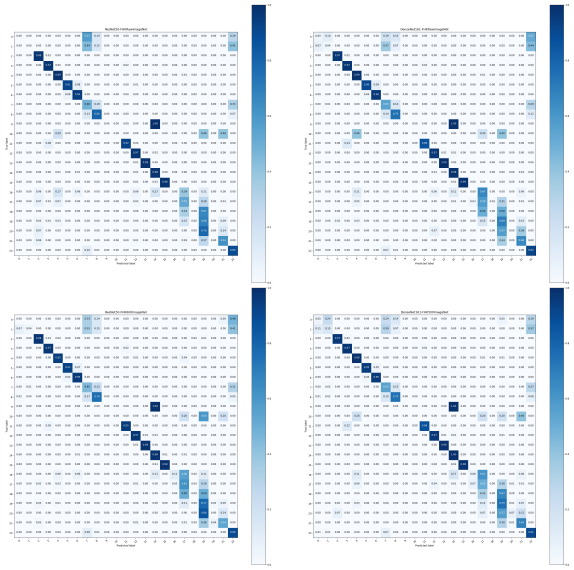


Fig. 2: [Best viewed in electronic version.] Confusion matrices of best performing raw configurations compared to their best performing augmented one on *Hyper-Kvasir* [29] dataset. 0: Barrett's, 1: Barrett's Short Segment, 2: BBPS-0-1, 3: BBPS-2-3, 4: Cecum, 5: Dyed Lifted Polyps, 6: Dyed Resection Margins, 7: Esophagitis a, 8: Esophagitis b-d, 9: Hemorrhoid, 10: Ileum, 11: Impacted Stool, 12: Polyps, 13: Pylorus, 14: Retroflex Rectum, 15: Retroflex Stomach, 16: Ulcerative Colitis grade 0-1, 17: Ulcerative Colitis grade 1, 18: Ulcerative Colitis grade 1-2, 19: Ulcerative Colitis grade 2, 20: Ulcerative Colitis grade 2-3, 21: Ulcerative Colitis grade 3, 22: Z-line

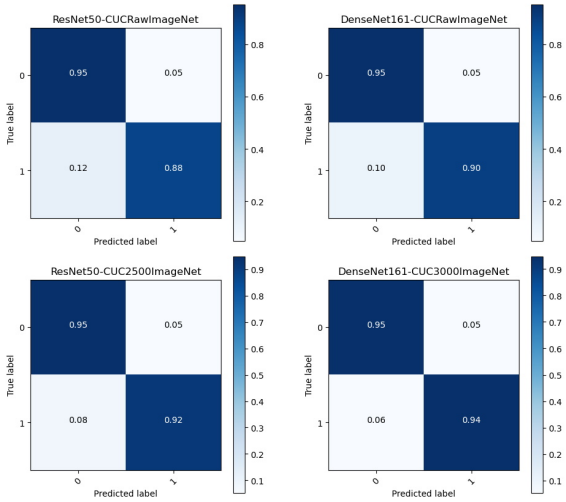


Fig. 3: [Best viewed in electronic version.] Confusion matrices of best performing raw configurations compared to their best performing augmented one on *Custom-UC*. 0: Non-pathological tract, 1: Pathological

augmentation strategy regarding the metric results. Improvements made in the various metrics show GAN's ability to provide more training information to the learner even with high FID levels.

Finally, the augmentation strategy chosen to deal with a large number of classes might present too much simplicity for a large multi-class problem. Indeed, the uniform addition of generated samples might hide the need for class-specific tweaking to properly augment each class as needed by the learner to understand class features correctly. These experiments let the possibility of additional work to provide the best-suited configuration for this specific dataset.

V. Conclusion and Future Works

In this work, we demonstrate a solution to deal with the lack of data in classification based on StyleGAN2-ADA [12, 13] and the freezing of the discriminator [17] to provide class-aware image generation. This method helped to extend existing datasets with low amounts of data and increase performance accuracy by using GAN state-of-the-art works and unlabeled data. Performance gains demonstrated through several classification metrics ascertain the value of the proposed method on *Hyper-Kvasir* [29] and our custom subset. This is observed by the results obtained on our *Custom-UC* dataset, which aimed to present our method's performance on this current classification challenge for endoscopic imaging. The

method seems, however, to be unable to adequately perform when only fewer than 20 images are available.

The artificial generation has shown its ability to assist classifiers training and offers better convergence performance on pretrained and non-pretrained classifiers and with *Custom-UC*, and full *Hyper-Kvasir* [29] dataset. However, few sample classes are noticeably overfitted by the generator and decrease the improvements provided by our strategy.

REFERENCES

[1] A. J. de Groof, M. R. Struyvenberg, J. van der Putten, F. van der Sommen, K. N. Fockens, W. L. Curvers, S. Zinger, R. E. Pouw, and et al., "Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking," *Gastroenterology*, vol. 158, no. 4, pp. 915–929.e4, 2020.

[2] E. Bentley, D. Jenkins, F. Campbell, and B. Warren, "How could pathologists improve the initial diagnosis of colitis? Evidence from an international workshop," *Journal of Clinical Pathology*, vol. 55, no. 12, pp. 955–960, 2002.

[3] H. P. Bhambhvani and A. Zamora, "Deep learning enabled classification of Mayo endoscopic subscore in patients with ulcerative colitis," *European Journal of Gastroenterology & Hepatology*, vol. 33, no. 5, pp. 645–649, May 2021.

[4] P. Rutgeerts, W. J. Sandborn, B. G. Feagan, W. Reinisch, A. Olson, J. Johanns, S. Travers, D. Rachmilewitz, and et al., "Infliximab for Induction and Maintenance Therapy for Ulcerative Colitis," *New England Journal of Medicine*, vol. 353, no. 23, pp. 2462–2476, 2005.

[5] D. J. Niven, T. J. McCormick, S. E. Straus, B. R. Hemmelgarn, L. Jeffs, T. R. M. Barnes, and H. T. Stelfox, "Reproducibility of clinical research in critical care: a scoping review," *BMC Medicine*, vol. 16, Feb. 2018.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *NeurIPS*, vol. 25, 2012.

[7] E. M. Song, B. Park, C.-A. Ha, S. W. Hwang, S. H. Park, D.-H. Yang, B. D. Ye, S.-J. Myung, and et al., "Endoscopic diagnosis and treatment planning for colorectal polyps using a deep-learning model," *Scientific Reports*, vol. 10, no. 1, p. 30, 2020.

[8] L. Perez and J. Wang, "The Effectiveness of Data Augmentation in Image Classification using Deep Learning," *ArXiv*, 2017.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," in *NeurIPS*, 2014, p. 2672–2680.

[10] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *ICLR*, 2019.

[11] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *CVPR*. IEEE, 2019, pp. 4396–4405.

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *CVPR*. IEEE, 2020, pp. 8107–8116.

[13] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *NeurIPS*, vol. 33, 2020, pp. 12 104–12 114.

[14] K. Su, E. Zhou, X. Sun, C. Wang, D. Yu, and X. Luo, "Pre-trained StyleGAN Based Data Augmentation for Small Sample Brain CT Motion Artifacts Detection," in *Advanced Data Mining and Applications*, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 339–346.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *ICML*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 214–223.

[17] S. Mo, M. Cho, and J. Shin, "Freeze the discriminator: a simple baseline for fine-tuning gans," in *CVPRW*. IEEE, 2020.

[18] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka, and J. Yang, "Imbalanced data learning by minority class augmentation using capsule adversarial networks," *Neurocomputing*, 2020.

[19] X. Yi, E. Walia, and P. Babyn, "Generative Adversarial Network in Medical Imaging: A Review," *Medical Image Analysis*, vol. 58, p. 101552, Dec. 2019.

[20] T. Kanayama, Y. Kurose, K. Tanaka, K. Aida, S. Satoh, M. Kitsuregawa, and T. Harada, "Gastric Cancer Detection from Endoscopic Images Using Synthesis by GAN," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, ser. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 530–538.

[21] S. Motamed, P. Rogalla, and F. Khalvati, "Data augmentation using generative adversarial networks (GANs) for GAN-based detection of pneumonia and COVID-19 in chest x-ray images," *Informatics in Medicine Unlocked*, vol. 27, p. 100779, 2021.

[22] Y. Fu, M. Gong, G. Yang, and J. Zhou, "Data augmentation for cardiac magnetic resonance image using evolutionary GAN," in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Springer International Publishing, 2021, pp. 126–141.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NeurIPS. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6629–6640.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *CVPR*. IEEE, 2015.

[25] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "StyleGAN2 Distillation for Feed-forward Image Manipulation," in *EECV*. Springer International Publishing, 2020.

[26] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: Discovering interpretable gan controls," in *NeurIPS*, 2020.

[27] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[28] J. Lv, G. Li, X. Tong, W. Chen, J. Huang, C. Wang, and G. Yang, "Transfer learning enhanced generative adversarial networks for multi-channel mri reconstruction," *Computers in Biology and Medicine*, vol. 134, p. 104504, 2021.

[29] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, and et al., "HyperKvasir , a comprehensive multi-class image and video dataset for gastrointestinal endoscopy," *Scientific Data*, vol. 7, no. 1, p. 283, 2020.

[30] X. Dray, C. Li, J.-C. Saurin, F. Cholet, G. Rahmi, J. Le Mouel, C. Leandri, S. Lecleire, and et al., "CAD-CAP: une base de données française à vocation internationale, pour le développement et la validation d'outils de diagnostic assisté par ordinateur en vidéocapsule endoscopique du grêle," in *Actes des JFHOD 2018*, ser. Endoscopy, Thieme, Ed., vol. 50, Paris, France, 2018, p. 316.

[31] R. Vallée, A. D. Maissin, A. Coutrot, H. Mouchère, A. Bourreille, and N. Normand, "CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms," in *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*. International Society for Optics and Photonics, Feb. 2020.

[32] R. W. Stidham, W. Liu, S. Bishu, M. D. Rice, P. D. R. Higgins, J. Zhu, B. K. Nallamothu, and A. K. Waljee, "Performance of a Deep Learning Model vs Human Reviewers in Grading Endoscopic Disease Severity of Patients With Ulcerative Colitis," *JAMA network open*, vol. 2, no. 5, p. e193963, May 2019.

[33] E. J. Lai, A. H. Calderwood, G. Doros, O. K. Fix, and B. C. Jacobson, "The Boston Bowel Preparation Scale: A valid and reliable instrument for colonoscopy-oriented research," *Gastrointestinal endoscopy*, vol. 69, no. 3 Pt 2, pp. 620–625, Mar. 2009.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. IEEE, 2016, pp. 770–778.

[35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *CVPR*. IEEE, Jan. 2017.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[37] M. J. Chong and D. Forsyth, "Effectively unbiased FID and inception score and where to find them," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020.